Apostolos Antonacopoulos ·
Subhasis Chaudhuri · Rama Chellappa ·
Cheng-Lin Liu · Saumik Bhattacharya ·
Umapada Pal (Eds.)

LNCS 15321

Pattern Recognition

27th International Conference, ICPR 2024
Kolkata, India, December 1–5, 2024
Proceedings, Part XXI

21 Part XXI

ICPR
2024
INDIA

IAPR

Springer

Lecture Notes in Computer Science 15321

The series Lecture Notes in Computer Science (LNCS), including its subseries Lecture Notes in Artificial Intelligence (LNAI) and Lecture Notes in Bioinformatics (LNBI), has established itself as a medium for the publication of new developments in computer science and information technology research, teaching, and education.

LNCS enjoys close cooperation with the computer science R & D community, the series counts many renowned academics among its volume editors and paper authors, and collaborates with prestigious societies. Its mission is to serve this international community by providing an invaluable service, mainly focused on the publication of conference and workshop proceedings and postproceedings. LNCS commenced publication in 1973.

Apostolos Antonacopoulos ·
Subhasis Chaudhuri · Rama Chellappa ·
Cheng-Lin Liu · Saumik Bhattacharya ·
Umapada Pal
Editors

Pattern Recognition

27th International Conference, ICPR 2024
Kolkata, India, December 1–5, 2024
Proceedings, Part XXI

Springer

Editors

Apostolos Antonacopoulos
University of Salford
Salford, Lancashire, UK

Subhasis Chaudhuri
Indian Institute of Technology Bombay
Mumbai, Maharashtra, India

Rama Chellappa
Johns Hopkins University
Baltimore, MD, USA

Cheng-Lin Liu
Chinese Academy of Sciences
Beijing, China

Saumik Bhattacharya
IIT Kharagpur
Kharagpur, West Bengal, India

Umapada Pal
Indian Statistical Institute Kolkata
Kolkata, West Bengal, India

ISSN 0302-9743 ISSN 1611-3349 (electronic)
Lecture Notes in Computer Science
ISBN 978-3-031-78304-3 ISBN 978-3-031-78305-0 (eBook)
https://doi.org/10.1007/978-3-031-78305-0

This Springer imprint is published by the registered company Springer Nature Switzerland AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

President's Address

On behalf of the Executive Committee of the International Association for Pattern Recognition (IAPR), I am pleased to welcome you to the 27th International Conference on Pattern Recognition (ICPR 2024), the main scientific event of the IAPR.

After a completely digital ICPR in the middle of the COVID pandemic and the first hybrid version in 2022, we can now enjoy a fully back-to-normal ICPR this year. I look forward to hearing inspirational talks and keynotes, catching up with colleagues during the breaks and making new contacts in an informal way. At the same time, the conference landscape has changed. Hybrid meetings have made their entrance and will continue. It is exciting to experience how this will influence the conference. Planning for a major event like ICPR must take place over a period of several years. This means many decisions had to be made under a cloud of uncertainty, adding to the already large effort needed to produce a successful conference. It is with enormous gratitude, then, that we must thank the team of organizers for their hard work, flexibility, and creativity in organizing this ICPR. ICPR always provides a wonderful opportunity for the community to gather together. I can think of no better location than Kolkata to renew the bonds of our international research community.

Each ICPR is a bit different owing to the vision of its organizing committee. For 2024, the conference has six different tracks reflecting major themes in pattern recognition: Artificial Intelligence, Pattern Recognition and Machine Learning; Computer and Robot Vision; Image, Speech, Signal and Video Processing; Biometrics and Human Computer Interaction; Document Analysis and Recognition; and Biomedical Imaging and Bioinformatics. This reflects the richness of our field. ICPR 2024 also features two dozen workshops, seven tutorials, and 15 competitions; there is something for everyone. Many thanks to those who are leading these activities, which together add significant value to attending ICPR, whether in person or virtually. Because it is important for ICPR to be as accessible as possible to colleagues from all around the world, we are pleased that the IAPR, working with the ICPR organizers, is continuing our practice of awarding travel stipends to a number of early-career authors who demonstrate financial need. Last but not least, we are thankful to the Springer LNCS team for their effort to publish these proceedings.

Among the presentations from distinguished keynote speakers, we are looking forward to the three IAPR Prize Lectures at ICPR 2024. This year we honor the achievements of Tin Kam Ho (IBM Research) with the IAPR's most prestigious King-Sun Fu Prize "for pioneering contributions to multi-classifier systems, random decision forests, and data complexity analysis". The King-Sun Fu Prize is given in recognition of an outstanding technical contribution to the field of pattern recognition. It honors the memory of Professor King-Sun Fu who was instrumental in the founding of IAPR, served as its first president, and is widely recognized for his extensive contributions to the field of pattern recognition.

The Maria Petrou Prize is given to a living female scientist/engineer who has made substantial contributions to the field of Pattern Recognition and whose past contributions, current research activity and future potential may be regarded as a model to both aspiring and established researchers. It honours the memory of Professor Maria Petrou as a scientist of the first rank, and particularly her role as a pioneer for women researchers. This year, the Maria Petrou Prize is given to Guoying Zhao (University of Oulu), "for contributions to video analysis for facial micro-behavior recognition and remote bio-signal reading (RPPG) for heart rate analysis and face anti-spoofing".

The J.K. Aggarwal Prize is given to a young scientist who has brought a substantial contribution to a field that is relevant to the IAPR community and whose research work has had a major impact on the field. Professor Aggarwal is widely recognized for his extensive contributions to the field of pattern recognition and for his participation in IAPR's activities. This year, the J.K. Aggarwal Prize goes to Xiaolong Wang (UC San Diego) "for groundbreaking contributions to advancing visual representation learning, utilizing self-supervised and attention-based models to establish fundamental frameworks for creating versatile, general-purpose pattern recognition systems".

During the conference we will also recognize 21 new IAPR Fellows selected from a field of very strong candidates. In addition, a number of Best Scientific Paper and Best Student Paper awards will be presented, along with the Best Industry Related Paper Award and the Piero Zamperoni Best Student Paper Award. Congratulations to the recipients of these very well-deserved awards!

I would like to close by again thanking everyone involved in making ICPR 2024 a tremendous success; your hard work is deeply appreciated. These thanks extend to all who chaired the various aspects of the conference and the associated workshops, my ExCo colleagues, and the IAPR Standing and Technical Committees. Linda O'Gorman, the IAPR Secretariat, deserves special recognition for her experience, historical perspective, and attention to detail when it comes to supporting many of the IAPR's most important activities. Her tasks became so numerous that she recently got support from Carolyn Buckley (layout, newsletter), Ugur Halici (ICPR matters), and Rosemary Stramka (secretariat). The IAPR website got a completely new design. Ed Sobczak has taken care of our web presence for so many years already. A big thank you to all of you!

This is, of course, the 27th ICPR conference. Knowing that ICPR is organized every two years, and that the first conference in the series (1973!) pre-dated the formal founding of the IAPR by a few years, it is also exciting to consider that we are celebrating over 50 years of ICPR and at the same time approaching the official IAPR 50th anniversary in 2028: you'll get all information you need at ICPR 2024. In the meantime, I offer my thanks and my best wishes to all who are involved in supporting the IAPR throughout the world.

September 2024

Arjan Kuijper
President of the IAPR

Preface

It is our great pleasure to welcome you to the proceedings of the 27th International Conference on Pattern Recognition (ICPR 2024), held in Kolkata, India. The city, formerly known as ‘Calcutta’, is the home of the fabled Indian Statistical Institute (ISI), which has been at the forefront of statistical pattern recognition for almost a century. Concepts like the Mahalanobis distance, Bhattacharyya bound, Cramer–Rao bound, and Fisher–Rao metric were invented by pioneers associated with ISI. The first ICPR (called IJCPR then) was held in 1973, and the second in 1974. Subsequently, ICPR has been held every other year. The International Association for Pattern Recognition (IAPR) was founded in 1978 and became the sponsor of the ICPR series. Over the past 50 years, ICPR has attracted huge numbers of scientists, engineers and students from all over the world and contributed to advancing research, development and applications in pattern recognition technology.

ICPR 2024 was held at the Biswa Bangla Convention Centre, one of the largest such facilities in South Asia, situated just 7 kilometers from Kolkata Airport (CCU). According to ChatGPT “Kolkata is often called the ‘Cultural Capital of India’. The city has a deep connection to literature, music, theater, and art. It was home to Nobel laureate Rabindranath Tagore, and the Bengali film industry has produced globally renowned filmmakers like Satyajit Ray. The city boasts remarkable colonial architecture, with landmarks like Victoria Memorial, Howrah Bridge, and the Indian Museum (the oldest and largest museum in India). Kolkata’s streets are dotted with old mansions and buildings that tell stories of its colonial past. Walking through the city can feel like stepping back into a different era. Finally, Kolkata is also known for its street food.”

ICPR 2024 followed a two-round paper submission format. We received a total of 2135 papers (1501 papers in round-1 submissions, and 634 papers in round-2 submissions). Each paper, on average, received 2.84 reviews, in single-blind mode. For the first-round papers we had a rebuttal option available to authors.

In total, 945 papers (669 from round-1 and 276 from round-2) were accepted for presentation, resulting in an acceptance rate of 44.26%, which is consistent with previous ICPR events. At ICPR 2024 the papers were categorized into six tracks: Artificial Intelligence, Machine Learning for Pattern Analysis; Computer Vision and Robotic Perception; Image, Video, Speech, and Signal Analysis; Biometrics and Human-Machine Interaction; Document and Media Analysis; and Biomedical Image Analysis and Informatics.

The main conference ran over December 2–5, 2024. The main program included the presentation of 188 oral papers (19.89% of the accepted papers), 757 poster papers and 12 competition papers (out of 15 submitted). A total 10 oral sessions were held concurrently in four meeting rooms with a total of 40 oral sessions. In total 24 workshops and 7 tutorials were held on December 1, 2024.

The plenary sessions included three prize lectures and three invited presentations. The prize lectures were delivered by Tin Kam Ho (IBM Research, USA; King Sun

Fu Prize winner), Xiaolong Wang (University of California, San Diego, USA; J.K. Aggarwal Prize winner), and Guoying Zhao (University of Oulu, Finland; Maria Petrou Prize winner). The invited speakers were Timothy Hospedales (University of Edinburgh, UK), Venu Govindaraju (University at Buffalo, USA), and Shuicheng Yan (Skywork AI, Singapore).

Several best paper awards were presented in ICPR: the Piero Zamperoni Award for the best paper authored by a student, the BIRPA Best Industry Related Paper Award, and the Best Paper Awards and Best Student Paper Awards for each of the six tracks of ICPR 2024.

The organization of such a large conference would not be possible without the help of many volunteers. Our special gratitude goes to the Program Chairs (Apostolos Antonacopoulos, Subhasis Chaudhuri, Rama Chellappa and Cheng-Lin Liu), for their leadership in organizing the program. Thanks to our Publication Chairs (Ananda S. Chowdhury and Wataru Ohyama) for handling the overwhelming workload of publishing the conference proceedings. We also thank our Competition Chairs (Richard Zanibbi, Lianwen Jin and Laurence Likforman-Sulem) for arranging 12 important competitions as part of ICPR 2024. We are thankful to our Workshop Chairs (P. Shivakumara, Stephanie Schuckers, Jean-Marc Ogier and Prabir Bhattacharya) and Tutorial Chairs (B.B. Chaudhuri, Michael R. Jenkin and Guoying Zhao) for arranging the workshops and tutorials on emerging topics. ICPR 2024, for the first time, held a Doctoral Consortium. We would like to thank our Doctoral Consortium Chairs (Véronique Eglin, Dan Lopresti and Mayank Vatsa) for organizing it.

Thanks go to the Track Chairs and the meta reviewers who devoted significant time to the review process and preparation of the program. We also sincerely thank the reviewers who provided valuable feedback to the authors.

Finally, we acknowledge the work of other conference committee members, like the Organizing Chairs and Organizing Committee Members, Finance Chairs, Award Chair, Sponsorship Chairs, and Exhibition and Demonstration Chairs, Visa Chair, Publicity Chairs, and Women in ICPR Chairs, whose efforts made this event successful. We also thank our event manager Alpcord Network for their help.

We hope that all the participants found the technical program informative and enjoyed the sights, culture and cuisine of Kolkata.

October 2024

Umapada Pal
Josef Kittler
Anil Jain

Organization

General Chairs

Umapada Pal	Indian Statistical Institute, Kolkata, India
Josef Kittler	University of Surrey, UK
Anil Jain	Michigan State University, USA

Program Chairs

Apostolos Antonacopoulos	University of Salford, UK
Subhasis Chaudhuri	Indian Institute of Technology, Bombay, India
Rama Chellappa	Johns Hopkins University, USA
Cheng-Lin Liu	Institute of Automation, Chinese Academy of Sciences, China

Publication Chairs

Ananda S. Chowdhury	Jadavpur University, India
Wataru Ohyama	Tokyo Denki University, Japan

Competition Chairs

Richard Zanibbi	Rochester Institute of Technology, USA
Lianwen Jin	South China University of Technology, China
Laurence Likforman-Sulem	Télécom Paris, France

Workshop Chairs

P. Shivakumara	University of Salford, UK
Stephanie Schuckers	Clarkson University, USA
Jean-Marc Ogier	Université de la Rochelle, France
Prabir Bhattacharya	Concordia University, Canada

Tutorial Chairs

B. B. Chaudhuri	Indian Statistical Institute, Kolkata, India
Michael R. Jenkin	York University, Canada
Guoying Zhao	University of Oulu, Finland

Doctoral Consortium Chairs

Véronique Eglin	CNRS, France
Daniel P. Lopresti	Lehigh University, USA
Mayank Vatsa	Indian Institute of Technology, Jodhpur, India

Organizing Chairs

Saumik Bhattacharya	Indian Institute of Technology, Kharagpur, India
Palash Ghosal	Sikkim Manipal University, India

Organizing Committee

Santanu Phadikar	West Bengal University of Technology, India
SK Md Obaidullah	Aliah University, India
Sayantari Ghosh	National Institute of Technology Durgapur, India
Himadri Mukherjee	West Bengal State University, India
Nilamadhaba Tripathy	Clarivate Analytics, USA
Chayan Halder	West Bengal State University, India
Shibaprasad Sen	Techno Main Salt Lake, India

Finance Chairs

Kaushik Roy	West Bengal State University, India
Michael Blumenstein	University of Technology Sydney, Australia

Awards Committee Chair

Arpan Pal	Tata Consultancy Services, India

Sponsorship Chairs

P. J. Narayanan	Indian Institute of Technology, Hyderabad, India
Yasushi Yagi	Osaka University, Japan
Venu Govindaraju	University at Buffalo, USA
Alberto Bel Bimbo	Università di Firenze, Italy

Exhibition and Demonstration Chairs

Arjun Jain	FastCode AI, India
Agnimitra Biswas	National Institute of Technology, Silchar, India

International Liaison, Visa Chair

Balasubramanian Raman	Indian Institute of Technology, Roorkee, India

Publicity Chairs

Dipti Prasad Mukherjee	Indian Statistical Institute, Kolkata, India
Bob Fisher	University of Edinburgh, UK
Xiaojun Wu	Jiangnan University, China

Women in ICPR Chairs

Ingela Nystrom	Uppsala University, Sweden
Alexandra B. Albu	University of Victoria, Canada
Jing Dong	Institute of Automation, Chinese Academy of Sciences, China
Sarbani Palit	Indian Statistical Institute, Kolkata, India

Event Manager

Alpcord Network

Track Chairs – Artificial Intelligence, Machine Learning for Pattern Analysis

Larry O'Gorman	Nokia Bell Labs, USA
Dacheng Tao	University of Sydney, Australia
Petia Radeva	University of Barcelona, Spain
Susmita Mitra	Indian Statistical Institute, Kolkata, India
Jiliang Tang	Michigan State University, USA

Track Chairs – Computer and Robot Vision

C. V. Jawahar	International Institute of Information Technology (IIIT), Hyderabad, India
João Paulo Papa	São Paulo State University, Brazil
Maja Pantic	Imperial College London, UK
Gang Hua	Dolby Laboratories, USA
Junwei Han	Northwestern Polytechnical University, China

Track Chairs – Image, Speech, Signal and Video Processing

P. K. Biswas	Indian Institute of Technology, Kharagpur, India
Shang-Hong Lai	National Tsing Hua University, Taiwan
Hugo Jair Escalante	INAOE, CINVESTAV, Mexico
Sergio Escalera	Universitat de Barcelona, Spain
Prem Natarajan	University of Southern California, USA

Track Chairs – Biometrics and Human Computer Interaction

Richa Singh	Indian Institute of Technology, Jodhpur, India
Massimo Tistarelli	University of Sassari, Italy
Vishal Patel	Johns Hopkins University, USA
Wei-Shi Zheng	Sun Yat-sen University, China
Jian Wang	Snap, USA

Track Chairs – Document Analysis and Recognition

Xiang Bai	Huazhong University of Science and Technology, China
David Doermann	University at Buffalo, USA
Josep Llados	Universitat Autònoma de Barcelona, Spain
Mita Nasipuri	Jadavpur University, India

Track Chairs – Biomedical Imaging and Bioinformatics

Jayanta Mukhopadhyay	Indian Institute of Technology, Kharagpur, India
Xiaoyi Jiang	Universität Münster, Germany
Seong-Whan Lee	Korea University, Korea

Metareviewers (Conference Papers and Competition Papers)

Wael Abd-Almageed	University of Southern California, USA
Maya Aghaei	NHL Stenden University, Netherlands
Alireza Alaei	Southern Cross University, Australia
Rajagopalan N. Ambasamudram	Indian Institute of Technology, Madras, India
Suyash P. Awate	Indian Institute of Technology, Bombay, India
Inci M. Baytas	Bogazici University, Turkey
Aparna Bharati	Lehigh University, USA
Brojeshwar Bhowmick	Tata Consultancy Services, India
Jean-Christophe Burie	University of La Rochelle, France
Gustavo Carneiro	University of Surrey, UK
Chee Seng Chan	Universiti Malaya, Malaysia
Sumohana S. Channappayya	Indian Institute of Technology, Hyderabad, India
Dongdong Chen	Microsoft, USA
Shengyong Chen	Tianjin University of Technology, China
Jun Cheng	Institute for Infocomm Research, A*STAR, Singapore
Albert Clapés	University of Barcelona, Spain
Oscar Dalmau	Center for Research in Mathematics, Mexico

Tyler Derr	Vanderbilt University, USA
Abhinav Dhall	Indian Institute of Technology, Ropar, India
Bo Du	Wuhan University, China
Yuxuan Du	University of Sydney, Australia
Ayman S. El-Baz	University of Louisville, USA
Francisco Escolano	University of Alicante, Spain
Siamac Fazli	Nazarbayev University, Kazakhstan
Jianjiang Feng	Tsinghua University, China
Gernot A. Fink	TU Dortmund University, Germany
Alicia Fornes	CVC, Spain
Junbin Gao	University of Sydney, Australia
Yan Gao	Amazon, USA
Yongsheng Gao	Griffith University, Australia
Caren Han	University of Melbourne, Australia
Ran He	Institute of Automation, Chinese Academy of Sciences, China
Tin Kam Ho	IBM, USA
Di Huang	Beihang University, China
Kaizhu Huang	Duke Kunshan University, China
Donato Impedovo	University of Bari, Italy
Julio Jacques	University of Barcelona and Computer Vision Center, Spain
Lianwen Jin	South China University of Technology, China
Wei Jin	Emory University, USA
Danilo Samuel Jodas	São Paulo State University, Brazil
Manjunath V. Joshi	DA-IICT, India
Jayashree Kalpathy-Cramer	Massachusetts General Hospital, USA
Dimosthenis Karatzas	Computer Vision Centre, Spain
Hamid Karimi	Utah State University, USA
Baiying Lei	Shenzhen University, China
Guoqi Li	Chinese Academy of Sciences, and Peng Cheng Lab, China
Laurence Likforman-Sulem	Institut Polytechnique de Paris/Télécom Paris, France
Aishan Liu	Beihang University, China
Bo Liu	Bytedance, USA
Chen Liu	Clarkson University, USA
Cheng-Lin Liu	Institute of Automation, Chinese Academy of Sciences, China
Hongmin Liu	University of Science and Technology Beijing, China
Hui Liu	Michigan State University, USA

Jing Liu	Institute of Automation, Chinese Academy of Sciences, China
Li Liu	University of Oulu, Finland
Qingshan Liu	Nanjing University of Posts and Telecommunications, China
Adrian P. Lopez-Monroy	Centro de Investigacion en Matematicas AC, Mexico
Daniel P. Lopresti	Lehigh University, USA
Shijian Lu	Nanyang Technological University, Singapore
Yong Luo	Wuhan University, China
Andreas K. Maier	FAU Erlangen-Nuremberg, Germany
Davide Maltoni	University of Bologna, Italy
Hong Man	Stevens Institute of Technology, USA
Lingtong Min	Northwestern Polytechnical University, China
Paolo Napoletano	University of Milano-Bicocca, Italy
Kamal Nasrollahi	Milestone Systems, Aalborg University, Denmark
Marcos Ortega	University of A Coruña, Spain
Shivakumara Palaiahnakote	University of Salford, UK
P. Jonathon Phillips	NIST, USA
Filiberto Pla	University Jaume I, Spain
Ajit Rajwade	Indian Institute of Technology, Bombay, India
Shanmuganathan Raman	Indian Institute of Technology, Gandhinagar, India
Imran Razzak	UNSW, Australia
Beatriz Remeseiro	University of Oviedo, Spain
Gustavo Rohde	University of Virginia, USA
Partha Pratim Roy	Indian Institute of Technology, Roorkee, India
Sanjoy K. Saha	Jadavpur University, India
Joan Andreu Sánchez	Universitat Politècnica de València, Spain
Claudio F. Santos	UFSCar, Brazil
Shin'ichi Satoh	National Institute of Informatics, Japan
Stephanie Schuckers	Clarkson University, USA
Srirangaraj Setlur	University at Buffalo, SUNY, USA
Debdoot Sheet	Indian Institute of Technology, Kharagpur, India
Jun Shen	University of Wollongong, Australia
Li Shen	JD Explore Academy, China
Chen Shengyong	Zhejiang University of technology and Tianjin University of Technology, China
Andy Song	RMIT University, Australia
Akihiro Sugimoto	National Institute of Informatics, Japan
Qianru Sun	Singapore Management University, Singapore
Arijit Sur	Indian Institute of Technology, Guwahati, India
Estefania Talavera	University of Twente, Netherlands

Wei Tang	University of Illinois at Chicago, USA
Joao M. Tavares	Universidade do Porto, Portugal
Jun Wan	NLPR, CASIA, China
Le Wang	Xi'an Jiaotong University, China
Lei Wang	Australian National University, Australia
Xiaoyang Wang	Tencent AI Lab, USA
Xinggang Wang	Huazhong University of Science and Technology, China
Xiao-Jun Wu	Jiangnan University, China
Yiding Yang	Bytedance, China
Xiwen Yao	Northwestern Polytechnical University, China
Xu-Cheng Yin	University of Science and Technology Beijing, China
Baosheng Yu	University of Sydney, Australia
Shiqi Yu	Southern University of Science and Technology, China
Xin Yuan	Westlake University, China
Yibing Zhan	JD Explore Academy, China
Jing Zhang	University of Sydney, Australia
Lefei Zhang	Wuhan University, China
Min-Ling Zhang	Southeast University, China
Wenbin Zhang	Florida International University, USA
Jiahuan Zhou	Peking University, China
Sanping Zhou	Xi'an Jiaotong University, China
Tianyi Zhou	University of Maryland, USA
Lei Zhu	Shandong Normal University, China
Pengfei Zhu	Tianjin University, China
Wangmeng Zuo	Harbin Institute of Technology, China

Reviewers (Competition Papers)

Liangcai Gao
Mingxin Huang
Lei Kang
Wenhui Liao
Yuliang Liu
Yongxin Shi
Da-Han Wang
Yang Xue
Wentao Yang
Jiaxin Zhang
Yiwu Zhong

Reviewers (Conference Papers)

Aakanksha Aakanksha
Aayush Singla
Abdul Muqeet
Abhay Yadav
Abhijeet Vijay Nandedkar
Abhimanyu Sahu
Abhinav Rajvanshi
Abhisek Ray
Abhishek Shrivastava
Abhra Chaudhuri
Aditi Roy
Adriano Simonetto
Adrien Maglo
Ahmed Abdulkadir
Ahmed Boudissa
Ahmed Hamdi
Ahmed Rida Sekkat
Ahmed Sharafeldeen
Aiman Farooq
Aishwarya Venkataramanan
Ajay Kumar
Ajay Kumar Reddy Poreddy
Ajita Rattani
Ajoy Mondal
Akbar K.
Akbar Telikani
Akshay Agarwal
Akshit Jindal
Al Zadid Sultan Bin Habib
Albert Clapés
Alceu Britto
Alejandro Peña
Alessandro Ortis
Alessia Auriemma Citarella
Alexandre Stenger
Alexandros Sopasakis
Alexia Toumpa
Ali Khan
Alik Pramanick
Alireza Alaei
Alper Yilmaz
Aman Verma
Amit Bhardwaj
Amit More
Amit Nandedkar
Amitava Chatterjee
Amos L. Abbott
Amrita Mohan
Anand Mishra
Ananda S. Chowdhury
Anastasia Zakharova
Anastasios L. Kesidis
Andras Horvath
Andre Gustavo Hochuli
André P. Kelm
Andre Wyzykowski
Andrea Bottino
Andrea Lagorio
Andrea Torsello
Andreas Fischer
Andreas K. Maier
Andreu Girbau Xalabarder
Andrew Beng Jin Teoh
Andrew Shin
Andy J. Ma
Aneesh S. Chivukula
Ángela Casado-García
Anh Quoc Nguyen
Anindya Sen
Anirban Saha
Anjali Gautam
Ankan Bhattacharyya
Ankit Jha
Anna Scius-Bertrand
Annalisa Franco
Antoine Doucet
Antonino Staiano
Antonio Fernández
Antonio Parziale
Anu Singha
Anustup Choudhury
Anwesan Pal
Anwesha Sengupta
Archisman Adhikary
Arjan Kuijper
Arnab Kumar Das

Arnav Bhavsar
Arnav Varma
Arpita Dutta
Arshad Jamal
Artur Jordao
Arunkumar Chinnaswamy
Aryan Jadon
Aryaz Baradarani
Ashima Anand
Ashis Dhara
Ashish Phophalia
Ashok K. Bhateja
Ashutosh Vaish
Ashwani Kumar
Asifuzzaman Lasker
Atefeh Khoshkhahtinat
Athira Nambiar
Attilio Fiandrotti
Avandra S. Hemachandra
Avik Hati
Avinash Sharma
B. H. Shekar
B. Uma Shankar
Bala Krishna Thunakala
Balaji Tk
Balázs Pálffy
Banafsheh Adami
Bang-Dang Pham
Baochang Zhang
Baodi Liu
Bashirul Azam Biswas
Beiduo Chen
Benedikt Kottler
Beomseok Oh
Berkay Aydin
Berlin S. Shaheema
Bertrand Kerautret
Bettina Finzel
Bhavana Singh
Bibhas C. Dhara
Bilge Gunsel
Bin Chen
Bin Li
Bin Liu
Bin Yao
Bin-Bin Jia
Binbin Yong
Bindita Chaudhuri
Bindu Madhavi Tummala
Binh M. Le
Bi-Ru Dai
Bo Huang
Bo Jiang
Bob Zhang
Bowen Liu
Bowen Zhang
Boyang Zhang
Boyu Diao
Boyun Li
Brian M. Sadler
Bruce A. Maxwell
Bryan Bo Cao
Buddhika L. Semage
Bushra Jalil
Byeong-Seok Shin
Byung-Gyu Kim
Caihua Liu
Cairong Zhao
Camille Kurtz
Carlos A. Caetano
Carlos D. Martã-Nez-Hinarejos
Ce Wang
Cevahir Cigla
Chakravarthy Bhagvati
Chandrakanth Vipparla
Changchun Zhang
Changde Du
Changkun Ye
Changxu Cheng
Chao Fan
Chao Guo
Chao Qu
Chao Wen
Chayan Halder
Che-Jui Chang
Chen Feng
Chenan Wang
Cheng Yu
Chenghao Qian
Cheng-Lin Liu

Chengxu Liu
Chenru Jiang
Chensheng Peng
Chetan Ralekar
Chih-Wei Lin
Chih-Yi Chiu
Chinmay Sahu
Chintan Patel
Chintan Shah
Chiranjoy Chattopadhyay
Chong Wang
Choudhary Shyam Prakash
Christophe Charrier
Christos Smailis
Chuanwei Zhou
Chun-Ming Tsai
Chunpeng Wang
Ciro Russo
Claudio De Stefano
Claudio F. Santos
Claudio Marrocco
Connor Levenson
Constantine Dovrolis
Constantine Kotropoulos
Dai Shi
Dakshina Ranjan Kisku
Dan Anitei
Dandan Zhu
Daniela Pamplona
Danli Wang
Danqing Huang
Daoan Zhang
Daqing Hou
David A. Clausi
David Freire Obregon
David Münch
David Pujol Perich
Davide Marelli
De Zhang
Debalina Barik
Debapriya Roy (Kundu)
Debashis Das
Debashis Das Chakladar
Debi Prosad Dogra
Debraj D. Basu
Decheng Liu
Deen Dayal Mohan
Deep A. Patel
Deepak Kumar
Dengpan Liu
Denis Coquenet
Désiré Sidibé
Devesh Walawalkar
Dewan Md. Farid
Di Ming
Di Qiu
Di Yuan
Dian Jia
Dianmo Sheng
Diego Thomas
Diganta Saha
Dimitri Bulatov
Dimpy Varshni
Dingcheng Yang
Dipanjan Das
Dipanjyoti Paul
Divya Biligere Shivanna
Divya Saxena
Divya Sharma
Dmitrii Matveichev
Dmitry Minskiy
Dmitry V. Sorokin
Dong Zhang
Donghua Wang
Donglin Zhang
Dongming Wu
Dongqiangzi Ye
Dongqing Zou
Dongrui Liu
Dongyang Zhang
Dongzhan Zhou
Douglas Rodrigues
Duarte Folgado
Duc Minh Vo
Duoxuan Pei
Durai Arun Pannir Selvam
Durga Bhavani S.
Eckart Michaelsen
Elena Goyanes
Élodie Puybareau

Emanuele Vivoli
Emna Ghorbel
Enrique Naredo
Enyu Cai
Eric Patterson
Ernest Valveny
Eva Blanco-Mallo
Eva Breznik
Evangelos Sartinas
Fabio Solari
Fabiola De Marco
Fan Wang
Fangda Li
Fangyuan Lei
Fangzhou Lin
Fangzhou Luo
Fares Bougourzi
Farman Ali
Fatiha Mokdad
Fei Shen
Fei Teng
Fei Zhu
Feiyan Hu
Felipe Gomes Oliveira
Feng Li
Fengbei Liu
Fenghua Zhu
Fillipe D. M. De Souza
Flavio Piccoli
Flavio Prieto
Florian Kleber
Francesc Serratosa
Francesco Bianconi
Francesco Castro
Francesco Ponzio
Francisco Javier Hernández López
Frédéric Rayar
Furkan Osman Kar
Fushuo Huo
Fuxiao Liu
Fu-Zhao Ou
Gabriel Turinici
Gabrielle Flood
Gajjala Viswanatha Reddy
Gaku Nakano
Galal Binamakhashen
Ganesh Krishnasamy
Gang Pan
Gangyan Zeng
Gani Rahmon
Gaurav Harit
Gennaro Vessio
Genoveffa Tortora
George Azzopardi
Gerard Ortega
Gerardo E. Altamirano-Gomez
Gernot A. Fink
Gibran Benitez-Garcia
Gil Ben-Artzi
Gilbert Lim
Giorgia Minello
Giorgio Fumera
Giovanna Castellano
Giovanni Puglisi
Giulia Orrù
Giuliana Ramella
Gökçe Uludoğan
Gopi Ramena
Gorthi Rama Krishna Sai Subrahmanyam
Gourav Datta
Gowri Srinivasa
Gozde Sahin
Gregory Randall
Guanjie Huang
Guanjun Li
Guanwen Zhang
Guanyu Xu
Guanyu Yang
Guanzhou Ke
Guhnoo Yun
Guido Borghi
Guilherme Brandão Martins
Guillaume Caron
Guillaume Tochon
Guocai Du
Guohao Li
Guoqiang Zhong
Guorong Li
Guotao Li
Gurman Gill

Haechang Lee
Haichao Zhang
Haidong Xie
Haifeng Zhao
Haimei Zhao
Hainan Cui
Haixia Wang
Haiyan Guo
Hakime Ozturk
Hamid Kazemi
Han Gao
Hang Zou
Hanjia Lyu
Hanjoo Cho
Hanqing Zhao
Hanyuan Liu
Hanzhou Wu
Hao Li
Hao Meng
Hao Sun
Hao Wang
Hao Xing
Hao Zhao
Haoan Feng
Haodi Feng
Haofeng Li
Haoji Hu
Haojie Hao
Haojun Ai
Haopeng Zhang
Haoran Li
Haoran Wang
Haorui Ji
Haoxiang Ma
Haoyu Chen
Haoyue Shi
Harald Koestler
Harbinder Singh
Harris V. Georgiou
Hasan F. Ates
Hasan S. M. Al-Khaffaf
Hatef Otroshi Shahreza
Hebeizi Li
Heng Zhang
Hengli Wang
Hengyue Liu
Hertog Nugroho
Hieyong Jeong
Himadri Mukherjee
Hoai Ngo
Hoda Mohaghegh
Hong Liu
Hong Man
Hongcheng Wang
Hongjian Zhan
Hongxi Wei
Hongyu Hu
Hoseong Kim
Hossein Ebrahimnezhad
Hossein Malekmohamadi
Hrishav Bakul Barua
Hsueh-Yi Sean Lin
Hua Wei
Huafeng Li
Huali Xu
Huaming Chen
Huan Wang
Huang Chen
Huanran Chen
Hua-Wen Chang
Huawen Liu
Huayi Zhan
Hugo Jair Escalante
Hui Chen
Hui Li
Huichen Yang
Huiqiang Jiang
Huiyuan Yang
Huizi Yu
Hung T. Nguyen
Hyeongyu Kim
Hyeonjeong Park
Hyeonjun Lee
Hymalai Bello
Hyung-Gun Chi
Hyunsoo Kim
I-Chen Lin
Ik Hyun Lee
Ilan Shimshoni
Imad Eddine Toubal

Imran Sarker
Inderjot Singh Saggu
Indrani Mukherjee
Indranil Sur
Ines Rieger
Ioannis Pierros
Irina Rabaev
Ivan V. Medri
J. Rafid Siddiqui
Jacek Komorowski
Jacopo Bonato
Jacson Rodrigues Correia-Silva
Jaekoo Lee
Jaime Cardoso
Jakob Gawlikowski
Jakub Nalepa
James L. Wayman
Jan Čech
Jangho Lee
Jani Boutellier
Javier Gurrola-Ramos
Javier Lorenzo-Navarro
Jayasree Saha
Jean Lee
Jean Paul Barddal
Jean-Bernard Hayet
Jean-Philippe G. Tarel
Jean-Yves Ramel
Jenny Benois-Pineau
Jens Bayer
Jerin Geo James
Jesús Miguel García-Gorrostieta
Jia Qu
Jiahong Chen
Jiaji Wang
Jian Hou
Jian Liang
Jian Xu
Jian Zhu
Jianfeng Lu
Jianfeng Ren
Jiangfan Liu
Jianguo Wang
Jiangyan Yi
Jiangyong Duan
Jianhua Yang
Jianhua Zhang
Jianhui Chen
Jianjia Wang
Jianli Xiao
Jianqiang Xiao
Jianwu Wang
Jianxin Zhang
Jianxiong Gao
Jianxiong Zhou
Jianyu Wang
Jianzhong Wang
Jiaru Zhang
Jiashu Liao
Jiaxin Chen
Jiaxin Lu
Jiaxing Ye
Jiaxuan Chen
Jiaxuan Li
Jiayi He
Jiayin Lin
Jie Ou
Jiehua Zhang
Jiejie Zhao
Jignesh S. Bhatt
Jin Gao
Jin Hou
Jin Hu
Jin Shang
Jing Tian
Jing Yu Chen
Jingfeng Yao
Jinglun Feng
Jingtong Yue
Jingwei Guo
Jingwen Xu
Jingyuan Xia
Jingzhe Ma
Jinhong Wang
Jinjia Wang
Jinlai Zhang
Jinlong Fan
Jinming Su
Jinrong He
Jintao Huang

Jinwoo Ahn
Jinwoo Choi
Jinyang Liu
Jinyu Tian
Jionghao Lin
Jiuding Duan
Jiwei Shen
Jiyan Pan
Jiyoun Kim
João Papa
Johan Debayle
John Atanbori
John Wilson
John Zhang
Jónathan Heras
Joohi Chauhan
Jorge Calvo-Zaragoza
Jorge Figueroa
Jorma Laaksonen
José Joaquim De Moura Ramos
Jose Vicent
Joseph Damilola Akinyemi
Josiane Zerubia
Juan Wen
Judit Szücs
Juepeng Zheng
Juha Roning
Jumana H. Alsubhi
Jun Cheng
Jun Ni
Jun Wan
Junghyun Cho
Junjie Liang
Junjie Ye
Junlin Hu
Juntong Ni
Junxin Lu
Junxuan Li
Junyaup Kim
Junyeong Kim
Jürgen Seiler
Jushang Qiu
Juyang Weng
Jyostna Devi Bodapati
Jyoti Singh Kirar
Kai Jiang
Kaiqiang Song
Kalidas Yeturu
Kalle Åström
Kamalakar Vijay Thakare
Kang Gu
Kang Ma
Kanji Tanaka
Karthik Seemakurthy
Kaushik Roy
Kavisha Jayathunge
Kazuki Uehara
Ke Shi
Keigo Kimura
Keiji Yanai
Kelton A. P. Costa
Kenneth Camilleri
Kenny Davila
Ketan Atul Bapat
Ketan Kotwal
Kevin Desai
Keyu Long
Khadiga Mohamed Ali
Khakon Das
Khan Muhammad
Kilho Son
Kim-Ngan Nguyen
Kishan Kc
Kishor P. Upla
Klaas Dijkstra
Komal Bharti
Konstantinos Triaridis
Kostas Ioannidis
Koyel Ghosh
Kripabandhu Ghosh
Krishnendu Ghosh
Kshitij S. Jadhav
Kuan Yan
Kun Ding
Kun Xia
Kun Zeng
Kunal Banerjee
Kunal Biswas
Kunchi Li
Kurban Ubul

Lahiru N. Wijayasingha
Laines Schmalwasser
Lakshman Mahto
Lala Shakti Swarup Ray
Lale Akarun
Lan Yan
Lawrence Amadi
Lee Kang Il
Lei Fan
Lei Shi
Lei Wang
Leonardo Rossi
Lequan Lin
Levente Tamas
Li Bing
Li Li
Li Ma
Li Song
Lia Morra
Liang Xie
Liang Zhao
Lianwen Jin
Libing Zeng
Lidia Sánchez-González
Lidong Zeng
Lijun Li
Likang Wang
Lili Zhao
Lin Chen
Lin Huang
Linfei Wang
Ling Lo
Lingchen Meng
Lingheng Meng
Lingxiao Li
Lingzhong Fan
Liqi Yan
Liqiang Jing
Lisa Gutzeit
Liu Ziyi
Liushuai Shi
Liviu-Daniel Stefan
Liyuan Ma
Liyun Zhu
Lizuo Jin
Longteng Guo
Lorena Álvarez Rodríguez
Lorenzo Putzu
Lu Leng
Lu Pang
Lu Wang
Luan Pham
Luc Brun
Luca Guarnera
Luca Piano
Lucas Alexandre Ramos
Lucas Goncalves
Lucas M. Gago
Luigi Celona
Luis C. S. Afonso
Luis Gerardo De La Fraga
Luis S. Luevano
Luis Teixeira
Lunke Fei
M. Hassaballah
Maddimsetti Srinivas
Mahendran N.
Mahesh Mohan M. R.
Maiko Lie
Mainak Singha
Makoto Hirose
Malay Bhattacharyya
Mamadou Dian Bah
Man Yao
Manali J. Patel
Manav Prabhakar
Manikandan V. M.
Manish Bhatt
Manjunath Shantharamu
Manuel Curado
Manuel Günther
Manuel Marques
Marc A. Kastner
Marc Chaumont
Marc Cheong
Marc Lalonde
Marco Cotogni
Marcos C. Santana
Mario Molinara
Mariofanna Milanova

Markus Bauer
Marlon Becker
Mårten Wadenbäck
Martin G. Ljungqvist
Martin Kampel
Martina Pastorino
Marwan Torki
Masashi Nishiyama
Masayuki Tanaka
Massimo O. Spata
Matteo Ferrara
Matthew D. Dawkins
Matthew Gadd
Matthew S. Watson
Maura Pintor
Max Ehrlich
Maxim Popov
Mayukh Das
Md Baharul Islam
Md Sajid
Meghna Kapoor
Meghna P. Ayyar
Mei Wang
Meiqi Wu
Melissa L. Tijink
Meng Li
Meng Liu
Meng-Luen Wu
Mengnan Liu
Mengxi China Guo
Mengya Han
Michaël Clément
Michal Kawulok
Mickael Coustaty
Miguel Domingo
Milind G. Padalkar
Ming Liu
Ming Ma
Mingchen Feng
Mingde Yao
Minghao Li
Mingjie Sun
Ming-Kuang Daniel Wu
Mingle Xu
Mingyong Li
Mingyuan Jiu
Minh P. Nguyen
Minh Q. Tran
Minheng Ni
Minsu Kim
Minyi Zhao
Mirko Paolo Barbato
Mo Zhou
Modesto Castrillón-Santana
Mohamed Amine Mezghich
Mohamed Dahmane
Mohamed Elsharkawy
Mohamed Yousuf
Mohammad Hashemi
Mohammad Khalooei
Mohammad Khateri
Mohammad Mahdi Dehshibi
Mohammad Sadil Khan
Mohammed Mahmoud
Moises Diaz
Monalisha Mahapatra
Monidipa Das
Mostafa Kamali Tabrizi
Mridul Ghosh
Mrinal Kanti Bhowmik
Muchao Ye
Mugalodi Ramesha Rakesh
Muhammad Rameez Ur Rahman
Muhammad Suhaib Kanroo
Muming Zhao
Munender Varshney
Munsif Ali
Na Lv
Nader Karimi
Nagabhushan Somraj
Nakkwan Choi
Nakul Agarwal
Nan Pu
Nan Zhou
Nancy Mehta
Nand Kumar Yadav
Nandakishor Nandakishor
Nandyala Hemachandra
Nanfeng Jiang
Narayan Hegde

Narayan Ji Mishra
Narayan Vetrekar
Narendra D. Londhe
Nathalie Girard
Nati Ofir
Naval Kishore Mehta
Nazmul Shahadat
Neeti Narayan
Neha Bhargava
Nemanja Djuric
Newlin Shebiah R.
Ngo Ba Hung
Nhat-Tan Bui
Niaz Ahmad
Nick Theisen
Nicolas Passat
Nicolas Ragot
Nicolas Sidere
Nikolaos Mitianoudis
Nikolas Ebert
Nilah Ravi Nair
Nilesh A. Ahuja
Nilkanta Sahu
Nils Murrugarra-Llerena
Nina S. T. Hirata
Ninad Aithal
Ning Xu
Ningzhi Wang
Niraj Kumar
Nirmal S. Punjabi
Nisha Varghese
Norio Tagawa
Obaidullah Md Sk
Oguzhan Ulucan
Olfa Mechi
Oliver Tüselmann
Orazio Pontorno
Oriol Ramos Terrades
Osman Akin
Ouadi Beya
Ozge Mercanoglu Sincan
Pabitra Mitra
Padmanabha Reddy Y. C. A.
Palaash Agrawal
Palaiahnakote Shivakumara
Palash Ghosal
Pallav Dutta
Paolo Rota
Paramanand Chandramouli
Paria Mehrani
Parth Agrawal
Partha Basuchowdhuri
Patrick Horain
Pavan Kumar
Pavan Kumar Anasosalu Vasu
Pedro Castro
Peipei Li
Peipei Yang
Peisong Shen
Peiyu Li
Peng Li
Pengfei He
Pengrui Quan
Pengxin Zeng
Pengyu Yan
Peter Eisert
Petra Gomez-Krämer
Pierrick Bruneau
Ping Cao
Pingping Zhang
Pintu Kumar
Pooja Kumari
Pooja Sahani
Prabhu Prasad Dev
Pradeep Kumar
Pradeep Singh
Pranjal Sahu
Prasun Roy
Prateek Keserwani
Prateek Mittal
Praveen Kumar Chandaliya
Praveen Tirupattur
Pravin Nair
Preeti Gopal
Preety Singh
Prem Shanker Yadav
Prerana Mukherjee
Prerna A. Mishra
Prianka Dey
Priyanka Mudgal

Qc Kha Ng
Qi Li
Qi Ming
Qi Wang
Qi Zuo
Qian Li
Qiang Gan
Qiang He
Qiang Wu
Qiangqiang Zhou
Qianli Zhao
Qiansen Hong
Qiao Wang
Qidong Huang
Qihua Dong
Qin Yuke
Qing Guo
Qingbei Guo
Qingchao Zhang
Qingjie Liu
Qinhong Yang
Qiushi Shi
Qixiang Chen
Quan Gan
Quanlong Guan
Rachit Chhaya
Radu Tudor Ionescu
Rafal Zdunek
Raghavendra Ramachandra
Rahimul I. Mazumdar
Rahul Kumar Ray
Rajib Dutta
Rajib Ghosh
Rakesh Kumar
Rakesh Paul
Rama Chellappa
Rami O. Skaik
Ramon Aranda
Ran Wei
Ranga Raju Vatsavai
Ranganath Krishnan
Rasha Friji
Rashmi S.
Razaib Tariq
Rémi Giraud
René Schuster
Renlong Hang
Renrong Shao
Renu Sharma
Reza Sadeghian
Richard Zanibbi
Rimon Elias
Rishabh Shukla
Rita Delussu
Riya Verma
Robert J. Ravier
Robert Sablatnig
Robin Strand
Rocco Pietrini
Rocio Diaz Martin
Rocio Gonzalez-Diaz
Rohit Venkata Sai Dulam
Romain Giot
Romi Banerjee
Ru Wang
Ruben Machucho
Ruddy Théodose
Ruggero Pintus
Rui Deng
Rui P. Paiva
Rui Zhao
Ruifan Li
Ruigang Fu
Ruikun Li
Ruirui Li
Ruixiang Jiang
Ruowei Jiang
Rushi Lan
Rustam Zhumagambetov
S. Amutha
S. Divakar Bhat
Sagar Goyal
Sahar Siddiqui
Sahbi Bahroun
Sai Karthikeya Vemuri
Saibal Dutta
Saihui Hou
Sajad Ahmad Rather
Saksham Aggarwal
Sakthi U.

Salimeh Sekeh
Samar Bouazizi
Samia Boukir
Samir F. Harb
Samit Biswas
Samrat Mukhopadhyay
Samriddha Sanyal
Sandika Biswas
Sandip Purnapatra
Sanghyun Jo
Sangwoo Cho
Sanjay Kumar
Sankaran Iyer
Sanket Biswas
Santanu Roy
Santosh D. Pandure
Santosh Ku Behera
Santosh Nanabhau Palaskar
Santosh Prakash Chouhan
Sarah S. Alotaibi
Sasanka Katreddi
Sathyanarayanan N. Aakur
Saurabh Yadav
Sayan Rakshit
Scott McCloskey
Sebastian Bunda
Sejuti Rahman
Selim Aksoy
Sen Wang
Seraj A. Mostafa
Shanmuganathan Raman
Shao-Yuan Lo
Shaoyuan Xu
Sharia Arfin Tanim
Shehreen Azad
Sheng Wan
Shengdong Zhang
Shengwei Qin
Shenyuan Gao
Sherry X. Chen
Shibaprasad Sen
Shigeaki Namiki
Shiguang Liu
Shijie Ma
Shikun Li
Shinichiro Omachi
Shirley David
Shishir Shah
Shiv Ram Dubey
Shiva Baghel
Shivanand S. Gornale
Shogo Sato
Shotaro Miwa
Shreya Ghosh
Shreya Goyal
Shuai Su
Shuai Wang
Shuai Zheng
Shuaifeng Zhi
Shuang Qiu
Shuhei Tarashima
Shujing Lyu
Shuliang Wang
Shun Zhang
Shunming Li
Shunxin Wang
Shuping Zhao
Shuquan Ye
Shuwei Huo
Shuyue Lan
Shyi-Chyi Cheng
Si Chen
Siddarth Ravichandran
Sihan Chen
Siladittya Manna
Silambarasan Elkana Ebinazer
Simon Benaïchouche
Simon S. Woo
Simone Caldarella
Simone Milani
Simone Zini
Sina Lotfian
Sitao Luan
Sivaselvan B.
Siwei Li
Siwei Wang
Siwen Luo
Siyu Chen
Sk Aziz Ali
Sk Md Obaidullah

Sneha Shukla
Snehasis Banerjee
Snehasis Mukherjee
Snigdha Sen
Sofia Casarin
Soheila Farokhi
Soma Bandyopadhyay
Son Minh Nguyen
Son Xuan Ha
Sonal Kumar
Sonam Gupta
Sonam Nahar
Song Ouyang
Sotiris Kotsiantis
Souhaila Djaffal
Soumen Biswas
Soumen Sinha
Soumitri Chattopadhyay
Souvik Sengupta
Spiros Kostopoulos
Sreeraj Ramachandran
Sreya Banerjee
Srikanta Pal
Srinivas Arukonda
Stephane A. Guinard
Su O. Ruan
Subhadip Basu
Subhajit Paul
Subhankar Ghosh
Subhankar Mishra
Subhankar Roy
Subhash Chandra Pal
Subhayu Ghosh
Sudip Das
Sudipta Banerjee
Suhas Pillai
Sujit Das
Sukalpa Chanda
Sukhendu Das
Suklav Ghosh
Suman K. Ghosh
Suman Samui
Sumit Mishra
Sungho Suh
Sunny Gupta
Suraj Kumar Pandey
Surendrabikram Thapa
Suresh Sundaram
Sushil Bhattacharjee
Susmita Ghosh
Swakkhar Shatabda
Syed Ms Islam
Syed Tousiful Haque
Taegyeong Lee
Taihui Li
Takashi Shibata
Takeshi Oishi
Talha Ahmad Siddiqui
Tanguy Gernot
Tangwen Qian
Tanima Bhowmik
Tanpia Tasnim
Tao Dai
Tao Hu
Tao Sun
Taoran Yi
Tapan Shah
Taveena Lotey
Teng Huang
Tengqi Ye
Teresa Alarcon
Tetsuji Ogawa
Thanh Phuong Nguyen
Thanh Tuan Nguyen
Thattapon Surasak
Thibault Napolãon
Thierry Bouwmans
Thinh Truong Huynh Nguyen
Thomas De Min
Thomas E. K. Zielke
Thomas Swearingen
Tianatahina Jimmy Francky Randrianasoa
Tianheng Cheng
Tianjiao He
Tianyi Wei
Tianyuan Zhang
Tianyue Zheng
Tiecheng Song
Tilottama Goswami
Tim Büchner

Tim H. Langer
Tim Raven
Tingkai Liu
Tingting Yao
Tobias Meisen
Toby P. Breckon
Tong Chen
Tonghua Su
Tran Tuan Anh
Tri-Cong Pham
Trishna Saikia
Trung Quang Truong
Tuan T. Nguyen
Tuan Vo Van
Tushar Shinde
Ujjwal Karn
Ukrit Watchareeruetai
Uma Mudenagudi
Umarani Jayaraman
V. S. Malemath
Vallidevi Krishnamurthy
Ved Prakash
Venkata Krishna Kishore Kolli
Venkata R. Vavilthota
Venkatesh Thirugnana Sambandham
Verónica Maria Vasconcelos
Véronique Ve Eglin
Víctor E. Alonso-Pérez
Vinay Palakkode
Vinayak S. Nageli
Vincent J. Whannou De Dravo
Vincenzo Conti
Vincenzo Gattulli
Vineet Padmanabhan
Vishakha Pareek
Viswanath Gopalakrishnan
Vivek Singh Baghel
Vivekraj K.
Vladimir V. Arlazarov
Vu-Hoang Tran
W. Sylvia Lilly Jebarani
Wachirawit Ponghiran
Wafa Khlif
Wang An-Zhi
Wanli Xue
Wataru Ohyama
Wee Kheng Leow
Wei Chen
Wei Cheng
Wei Hua
Wei Lu
Wei Pan
Wei Tian
Wei Wang
Wei Wei
Wei Zhou
Weidi Liu
Weidong Yang
Weijun Tan
Weimin Lyu
Weinan Guan
Weining Wang
Weiqiang Wang
Weiwei Guo
Weixia Zhang
Wei-Xuan Bao
Weizhong Jiang
Wen Xie
Wenbin Qian
Wenbin Tian
Wenbin Wang
Wenbo Zheng
Wenhan Luo
Wenhao Wang
Wen-Hung Liao
Wenjie Li
Wenkui Yang
Wenwen Si
Wenwen Yu
Wenwen Zhang
Wenwu Yang
Wenxi Li
Wenxi Yue
Wenxue Cui
Wenzhuo Liu
Widhiyo Sudiyono
Willem Dijkstra
Wolfgang Fuhl
Xi Zhang
Xia Yuan

Xianda Zhang
Xiang Zhang
Xiangdong Su
Xiang-Ru Yu
Xiangtai Li
Xiangyu Xu
Xiao Guo
Xiao Hu
Xiao Wu
Xiao Yang
Xiaofeng Zhang
Xiaogang Du
Xiaoguang Zhao
Xiaoheng Jiang
Xiaohong Zhang
Xiaohua Huang
Xiaohua Li
Xiao-Hui Li
Xiaolong Sun
Xiaosong Li
Xiaotian Li
Xiaoting Wu
Xiaotong Luo
Xiaoyan Li
Xiaoyang Kang
Xiaoyi Dong
Xin Guo
Xin Lin
Xin Ma
Xinchi Zhou
Xingguang Zhang
Xingjian Leng
Xingpeng Zhang
Xingzheng Lyu
Xinjian Huang
Xinqi Fan
Xinqi Liu
Xinqiao Zhang
Xinrui Cui
Xizhan Gao
Xu Cao
Xu Ouyang
Xu Zhao
Xuan Shen
Xuan Zhou
Xuchen Li
Xuejing Lei
Xuelu Feng
Xueting Liu
Xuewei Li
Xueyi X. Wang
Xugong Qin
Xu-Qian Fan
Xuxu Liu
Xu-Yao Zhang
Yan Huang
Yan Li
Yan Wang
Yan Xia
Yan Zhuang
Yanan Li
Yanan Zhang
Yang Hou
Yang Jiao
Yang Liping
Yang Liu
Yang Qian
Yang Yang
Yang Zhao
Yangbin Chen
Yangfan Zhou
Yanhui Guo
Yanjia Huang
Yanjun Zhu
Yanming Zhang
Yanqing Shen
Yaoming Cai
Yaoxin Zhuo
Yaoyan Zheng
Yaping Zhang
Yaqian Liang
Yarong Feng
Yasmina Benmabrouk
Yasufumi Sakai
Yasutomo Kawanishi
Yazeed Alzahrani
Ye Du
Ye Duan
Yechao Zhang
Yeong-Jun Cho

Yi Huo
Yi Shi
Yi Yu
Yi Zhang
Yibo Liu
Yibo Wang
Yi-Chieh Wu
Yifan Chen
Yifei Huang
Yihao Ding
Yijie Tang
Yikun Bai
Yimin Wen
Yinan Yang
Yin-Dong Zheng
Yinfeng Yu
Ying Dai
Yingbo Li
Yiqiao Li
Yiqing Huang
Yisheng Lv
Yisong Xiao
Yite Wang
Yizhe Li
Yong Wang
Yonghao Dong
Yong-Hyuk Moon
Yongjie Li
Yongqian Li
Yongqiang Mao
Yongxu Liu
Yongyu Wang
Yongzhi Li
Youngha Hwang
Yousri Kessentini
Yu Wang
Yu Zhou
Yuan Tian
Yuan Zhang
Yuanbo Wen
Yuanxin Wang
Yubin Hu
Yubo Huang
Yuchen Ren
Yucheng Xing
Yuchong Yao
Yuecong Min
Yuewei Yang
Yufei Zhang
Yufeng Yin
Yugen Yi
Yuhang Ming
Yujia Zhang
Yujun Ma
Yukiko Kenmochi
Yun Hoyeoung
Yun Liu
Yunhe Feng
Yunxiao Shi
Yuru Wang
Yushun Tang
Yusuf Osmanlioglu
Yusuke Fujita
Yuta Nakashima
Yuwei Yang
Yuwu Lu
Yuxi Liu
Yuya Obinata
Yuyao Yan
Yuzhi Guo
Zaipeng Xie
Zander W. Blasingame
Zedong Wang
Zeliang Zhang
Zexin Ji
Zhanxiang Feng
Zhaofei Yu
Zhe Chen
Zhe Cui
Zhe Liu
Zhe Wang
Zhekun Luo
Zhen Yang
Zhenbo Li
Zhenchun Lei
Zhenfei Zhang
Zheng Liu
Zheng Wang
Zhengming Yu
Zhengyin Du

Zhengyun Cheng
Zhenshen Qu
Zhenwei Shi
Zhenzhong Kuang
Zhi Cai
Zhi Chen
Zhibo Chu
Zhicun Yin
Zhida Huang
Zhida Zhang
Zhifan Gao
Zhihang Ren
Zhihang Yuan
Zhihao Wang
Zhihua Xie
Zhihui Wang
Zhikang Zhang
Zhiming Zou
Zhiqi Shao
Zhiwei Dong
Zhiwei Qi
Zhixiang Wang
Zhixuan Li
Zhiyu Jiang
Zhiyuan Yan
Zhiyuan Yu
Zhiyuan Zhang
Zhong Chen
Zhongwei Teng
Zhongzhan Huang
Zhongzhi Yu
Zhuan Han
Zhuangzhuang Chen
Zhuo Liu
Zhuo Su
Zhuojun Zou
Zhuoyue Wang
Ziang Song
Zicheng Zhang
Zied Mnasri
Zifan Chen
Žiga Babnik
Zijing Chen
Zikai Zhang
Ziling Huang
Zilong Du
Ziqi Cai
Ziqi Zhou
Zi-Rui Wang
Zirui Zhou
Ziwen He
Ziyao Zeng
Ziyi Zhang
Ziyue Xiang
Zonglei Jing
Zongyi Xu

Contents – Part XXI

Long-Tailed Hashing with Wasserstein Quantization

Zujun Fu, Hanjiang Lai, and Yan Pan(✉)

School of Computer Science and Engineering, Sun Yat-Sen University, Guangzhou 510006, China
fuzj5@mail2.sysu.edu.cn, {laihanj3,panyan5}@mail.sysu.edu.cn

Abstract. Long-tailed hashing is to learn hash functions in unbalanced distribution datasets to represent images as binary hash codes for fast and accurate image retrieval. In contrast to balanced distribution datasets, unbalanced distributions are more common in the real world. However, Existing long-tailed hashing methods only focus on how to better learn from unbalanced datasets to improve performance, without giving good consideration to quantization error, which is very crucial in hash learning. In this paper, we propose a simple but efficient quantization method for long-tailed hashing. Specifically, to address the lack of samples in the tail classes, we take a uniform discrete distribution as the optimal target distribution. We use the Sliced Wasserstein distance as a measure of distribution distance. It makes good use of the discrete nature of hash functions and has low computational complexity. Then we formulate the optimization objective of the quantization error as minimizing the distance between the output of the learned hash function and this objective distribution, which can be added as an additional term of the loss function to existing long-tailed hashing methods. We conduct experiments on two long-tailed datasets, and the results show that our proposed method greatly improves the performance of existing long-tailed hashing methods.

Keywords: Image retrieval · Deep hashing · Long-tailed learning

1 Introduction

Hashing [21,24,27] is a widely used technique in image retrieval that can represent an image as a low-dimensional binary code using a function. It is characterized by fast retrieval, high accuracy, and low storage cost. However, current deep hashing methods are trained on ideally uniformly distributed data, which is uncommon in the real world. Therefore, it is necessary to learn from unbalanced long-tailed datasets. Long-tailed datasets are characterized by the fact that samples in the head classes account for most of the total sample volume, while samples in the tail classes are only a tiny number.

A. Antonacopoulos et al. (Eds.): ICPR 2024, LNCS 15321, pp. 1–14, 2025.
https://doi.org/10.1007/978-3-031-78305-0_1

There are already some hashing methods that deal with long-tailed distributions, such as LTHNet [5] and ACHNet [13]. LTHNet proposes a Dynamic Meta Embedding (DME) module to address the long-tailed distribution by transferring the knowledge from head classes to tail classes. ACHNet employs the attention mechanism and the contrastive learning of hash codes and proposes a Cross Attention Feature Enhanced (CAFE) module to mitigate the information loss caused by feature dimensionality reduction.

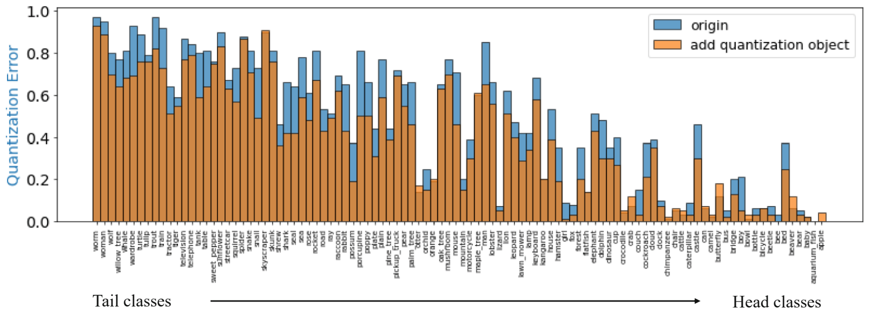


Fig. 1. Comparison of the average quantization error (angle between continuous codes and hash codes in radian) per class for the original long-tailed hashing method in the Cifar-100 dataset with the method after adding the quantization objective. The classes are listed in ascending order of sample size, with the head class on the left and the tail class on the right.

However, both methods only map the continuous output of the hash layer into binary code directly in the network using the *tanh* function, without considering the quantization error well. Quantization error is the loss of information that occurs when replacing a discrete function with a continuous function due to the difficulty of discrete optimization [27]. In long-tailed hashing, the quantization error is much more crucial. Fig. 1 shows the quantization error of each class in the long-tailed dataset. It can be found that the quantization error of the tail classes is much larger than that of the head classes, which will severely affect the retrieval performance. This is because the lack of samples in the tail classes makes it more difficult to learn a good hash function. Therefore, the quantization error generated by the long-tailed hashing during the learning process must be handled.

In this paper, we propose a simple but efficient quantization method for long-tailed hashing. First, we show that hash codes satisfying a uniform discrete distribution are optimal. In this case, the classes are uniformly distributed regardless of the number of samples. Meanwhile, data points of the same class are aggregated together, and different classes are separated from each other. Therefore, we set the quantization objective as minimizing the distance between the output distribution of the hash layer and this optimal distribution. Then, we introduce the Wasserstein-2 distance as a measure of the distance between two distributions. To estimate it more easily, we project the Wasserstein-2 distance to the one-dimensional case to solve for it, which is called the Sliced Wasserstein distance. Fig. 1 illustrates that the problem of quantization error in long-tailed hashing is mitigated with the addition of our proposed quantization objective, especially for the tail classes. Our main contributions are as follows:

- We use the Sliced Wasserstein distance as a measure of the distribution distance to set quantization objectives for long-tailed hashing.
- We reduce the quantization error by minimizing the quantization objective, and it can be combined with existing long-tailed hashing methods as an additional term in the loss function.
- We conduct comparative experiments on two long-tailed datasets, and the results show that the performance of existing long-tailed hashing methods is significantly improved, which validates the effectiveness of our proposed method.

2 Related work

2.1 Long-tailed learning

The phenomenon of long-tailed distributions is ubiquitous in reality, and it is essential to study how to learn from such data. We present existing methods in three aspects: data resampling, class rebalancing, and knowledge transfer. Data resampling attempts to resample an unbalanced dataset to force the data distribution to be balanced. This is done by repeatedly sampling the tail classes [4] or removing some data from the head classes [11], but leads to overfitting [4] and underfitting [11], respectively. Class rebalancing attempts to assign different weights to the head and tail classes in learning. This can be done by weighting the loss function [20] and the samples used [6]. The idea of knowledge transfer is that hidden knowledge can be shared between different classes. This can be done either by transferring the knowledge learned from the head class to the tail class [26] or by designing an additional module to enrich the representation of the samples in the head and tail classes [22].

2.2 Deep hashing

Deep hash learning obtains binary hash codes for fast retrieval by learning hash functions from a dataset. CNNH [28] is the earliest deep hashing method, which generates a binary hash code first and then learns the hash function from the generated hash code using a CNN. DPSH [18] is an end-to-end deep hashing model that optimizes the model using a loss function containing similarity information and quantization error. DSDH [17] exploits the label information by adding regularized linear regression loss to the optimization objective in addition to pairwise similarity information. HashNet [3] learns the exact binary hash code by using continuation approximation. CSQ [29] aggregates samples of the same class together and separates samples of different classes from each other by presetting the hash centers.

2.3 Wasserstein distance

The Wasserstein distance is derived from optimal transmission theory and is used to measure the difference between two distributions. It has been widely used in the field of computer vision, e.g., WGAN [2]. The advantage of the Wasserstein distance over the commonly used KL and JS divergence is that it can measure non-overlapping distributions [1]. However, estimating this distance is difficult [2]. [8] explored estimating the Wasserstein-2 distance via an optimal transmission formula, but it is computationally very expensive and requires exponentially large samples [7]. A variant called Sliced Wasserstein distance then has a polynomial complexity. It projects the data points in many random one-dimensional directions, called slices, and estimates the Wasserstein distance by taking the average distance of each data point on these slices [12].

3 Method

3.1 Problem formulation

For a given long-tailed dataset $x = \{x_1, x_2, ..., x_n\}$ and label $y = \{y_1, y_2, ..., y_n\}$, long-tailed hashing task first maps the data x into a real-valued feature vector f using a feature extractor E parameterized by θ:

$$f = E\left(x|\theta\right) \tag{1}$$

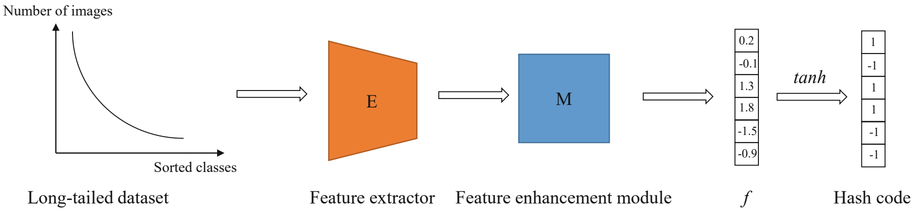


Fig. 2. The pipeline of general long-tailed hashing methods. The input long-tailed dataset goes through a feature extractor and feature enhancement module to output real-valued features, which are then mapped to binary hash codes using a *tanh* function.

To enhance the learning of tail classes, f usually goes through a feature enhancement module. The aim of long-tailed hash learning is to transform the feature f into a binary hash code consisting of $\{-1, +1\}$. Therefore, a hash layer is connected behind the feature enhancement module for the binary transformation $sgn(f)$. Due to the difficulty of discrete optimization, $sgn()$ needs to be replaced by a continuous function $h : f \rightarrow \{-1, +1\}^m$, where m is the length of the hash code. h is usually a $tanh()$ activation function. The pipeline of general long-tailed hashing methods is shown in Fig. 2.

Then, the optimization objective L of long-tailed hashing can be defined as:

$$\min_{\theta} L\left(h\left(f\right), y\right) \tag{2}$$

3.2 Code quantization

In hash learning, a quantization objective is also usually required. This is important to improve the retrieval performance of hashing methods because it reduces the quantization error. The quantization error is the loss of information that occurs when replacing a discrete function $sgn()$ with a continuous function $tanh()$ described in 3.1. A lower quantization error reduces the cases where data points belonging to the same class are assigned to hash codes with a larger Hamming distance. [27] suggests that this is very important in hash learning and helps to improve retrieval quality.

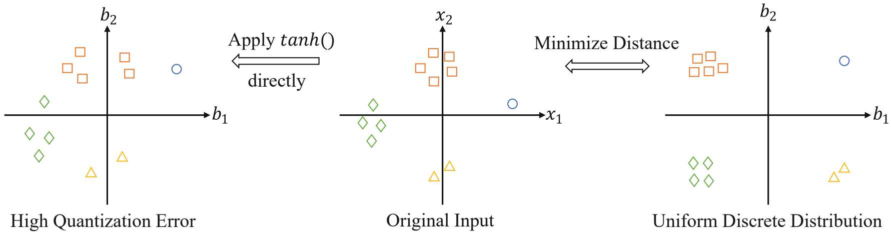


Fig. 3. Visual illustration of the optimal distribution (right) and quantization error of 2-bit hash codes (left). If the original input is mapped directly through the *tanh* function at the hash layer, a high quantization error occurs. Our goal is to minimize the distance between the original input and the uniform discrete distribution.

We consider the case of learning a 2-bit hash function. In Fig. 3, the learned hash function projects data from the four classes into a two-dimensional space as the original input to the hash layer. If these real-valued features are directly mapped to a hash code through the *tanh* function of the hash layer, a case with high quantization error will occur as shown in Fig. 3(left). Some data points of the same class near the boundary are assigned to two different hash codes, leading to a high false-negative rate and reducing the retrieval performance. Fig. 1 illustrates that this is more likely to happen with tail classes in long-tailed hashing.

Fig. 3(right) demonstrates the optimal hash distribution, in which data points of the same class are aggregated together and different classes are separated from each other, regardless of the number of samples. It is not affected by the lack of samples in the tail classes, ensuring a low quantization error. In addition, the data points are equally divided into four quadrants, which also ensures code balance. Therefore, we need to make the original distribution tend to the target distribution, i.e., minimize the distance between the two distributions.

3.3 Distributional distance

As described in 3.2, hash codes that follow a uniform discrete distribution are optimal for low quantization error. We denote this distribution by U. In this case, each bit of a hash code can be sampled independently and randomly as -1 or $+1$ with equal probability, which ensures coding balance. [9] shows that it is important and helps to reduce the average time complexity of retrieval. Therefore, we can use U as the target to minimize the difference between the output distribution of the hash layer and this optimal target distribution. We define the quantization objective L_q as follows:

$$L_q(h(f)) = D(h(f)||U) \tag{3}$$

where D denotes the distributional distance. However, minimizing D is difficult, especially if the density of the hash distribution cannot be estimated. Furthermore, the choice of a metric for the distribution distance D is also an issue to be considered.

The KL and JS divergences are well known and they are the most commonly used methods to measure the similarity of two distributions. However, using them to estimate D is not a good choice because they do not measure non-overlapping distributions and are computationally expensive. Therefore, we introduce the Wasserstein-2 distance to estimate D, which is more effective and computationally efficient than the KL and JS divergence. When using the Wasserstein-2 distance to measure the difference between two distributions, the distribution distance D is expressed as:

$$D\left(u,v\right) = \left(\underset{\gamma\in\varPi\left(u,v\right)}{inf}\int_{\left(a,b\right)\sim\gamma} p\left(a,b\right)\left\|a-b\right\|_2 dadb\right)^{\frac{1}{2}} \tag{4}$$

where $\varPi\left(u,v\right)$ is the set of all joint distributions $\gamma\left(a,b\right)$ with marginal distributions u and v, respectively. In simple words, $\gamma\left(a,b\right)$ denotes the "quality" of the transmission from a to b in order to transform the distribution u to the distribution v, and then the distance D is the "cost" of the optimal transmission plan. In Eq. (4), u and v represent the continuous hash distribution h and the discrete uniform distribution U, respectively. However, it is difficult to directly compute the infimum of Eq. (4) because the continuous hash distribution of the output of the network is not fixed or even unknown.

These problems can be avoided by using the Sliced Wasserstein distance. First, we let ρ_u and ρ_v denote the density functions of u and v respectively, and then the Wasserstein-2 distance W in the one-dimensional case is:

$$W\left(u,v\right) = \left(\int_0^1 \left\|F_u^{-1}\left(w\right) - F_v^{-1}\left(w\right)dw\right\|_2\right)^{\frac{1}{2}} \tag{5}$$

where $F_u\left(w\right) = \int_\infty^w \rho_u\left(\eta\right)d\eta$, $F_u\left(w\right) = \int_\infty^w \rho_u\left(\eta\right)d\eta$, which are the cumulative distribution functions of u and v, respectively. Then, we can utilize linear projection to approximate the Wasserstein-2 distance as a one-dimensional projection of these functions, which is the Sliced Wasserstein distance:

$$D\left(h\left(f\right),U\right)\approx\left(\frac{1}{L}\sum_{i=1}^{L}W\left(w_i^Th\left(f\right),w_i^TU\right)\right)^{\frac{1}{2}} \tag{6}$$

where $w_i^Th\left(f\right)$ and w_i^TU are one-dimensional projections in the direction of w_i (call slice) of the samples of the hash distribution h and the samples from U, and L is the number of slices. w_i is usually obtained by sampling on the unit sphere.

Furthermore, random projections can be avoided in the estimation of Sliced Wasserstein distance by choosing the direction that contains the information of the two distributions to be measured. In this way, Eq. (6) can be written as:

$$D\left(h\left(f\right),U\right)\approx\left(\frac{1}{m}\sum_{i=1}^{m}W\left(h\left(f\right)_{i,:},U_{i,:}\right)\right)^{\frac{1}{2}} \tag{7}$$

where $h\left(f\right)_{i,:}$ and $U_{i,:}$ are all one-dimensional samples of $h(f)$ and U in the i-dimensional direction respectively, m is the length of the hash code.

Finally, combining Eq. (2) and Eq. (3), we can define the quantization loss and add it to the optimization objective for long-tailed hashing:

$$\min_{\theta} L+\lambda L_q \tag{8}$$

4 Experiments

4.1 Datasets

Cifar-100 [14] Cifar-100 is a widely used data set in various fields. It contains 100 classes with a total of 60,000 images. Each class has a database of 500 images and a query set of 100 images. We use 50,000 of these images as a database, averaging 500 images per class, and the remaining 10,000 images as a query set, averaging 100 images per class. Here we randomly sample images from the database following Zipf's law, i.e. $N_i=N_1\times i^{-\mu}$, where N_i denotes the number of images from the i-th class and μ is an imbalance parameter. Then we build three unbalanced benchmarks according to (Imbalance Factor) IF = 1, IF = 50, and IF = 100 by choosing different μ. A larger IF means a more imbalanced training set while IF=1 means a balanced training set.

ImageNet-100 [3] ImageNet-100 is a 100-class subset of the original dataset. We construct an unbalanced dataset by randomly selecting 100 classes from all 1,000 classes. The database has a total of 130,000 images with 1,300 images per class and the query set has a total of 5,000 images with 50 images per class, where the training set follows Zipf's law and is randomly sampled from the database based on three different IF. For a fair comparison, we take only 100 images from each class as the training set when IF = 1.

The details of the two datasets after random sampling according to different IF are shown in Table 1 and Table 2.

Table 1. Details of the Cifar-100 dataset with different IF

IF	N_1	N_{100}	μ	N_{train}	$N_{database}$	N_{query}
1	500	500	0	50,000	50,000	10,000
50	500	10	0.83	3,732	50,000	10,000
100	500	5	0.99	2,598	50,000	10,000

Table 2. Details of the ImageNet-100 dataset with different IF

IF	N_1	N_{100}	μ	N_{train}	$N_{database}$	N_{query}
1	100	100	0	10,000	130,000	5,000
50	1300	26	0.845	9,437	130,000	5,000
100	1300	13	0.99	6,834	130,000	5,000

4.2 Implementation Details

We add the proposed quantization objective to two existing long-tailed hashing methods LTHNet[5] and ACHNet[13] with $\lambda = 0.1$, following the same hyper-parameter settings as the original methods. In addition, we also compare with four deep hashing methods, which are DPSH[18], HashNet[3], DSDH[17], and CSQ[29]. For a fair comparison, we used ResNet34 as a feature extractor for all methods. All the parameters of compared methods are optimized by a RMSprop algorithm with weight decay 5e-4. Learning rate is set to 1e-5 for Cifar-100 while 1e-6 for ImageNet-100, and these learning rates are updated by a cosine annealing schedule. Referring to previous methods, we use the mean average precision (MAP) as the evaluation metric and conduct the experiments under the settings of hash code lengths of 32bits, 64bits and 96bits, respectively. MAP has been used as the retrieval performance measure in almost all the learning to hash literature [10,15,16,19,25], models with higher MAP represent better performance.

4.3 Results and Analysis

We report the experiment results of various methods to learn 32-bit, 64-bit, and 96-bit hash functions on the Cifar-100 and ImageNet-100 datasets in Table 3 and Table 4, respectively. -Q denotes the addition of our proposed quantization objective to the original method, and bold values denote the performance improvement relative to the original method. It can be seen that after adding the quantization objective, the performance of LTHNet on the Cifar-100 dataset achieves an improvement of 0.96%-3.06%, and the performance of ACHNet achieves an improvement of 0.92%-2.67%. In the ImageNet-100 dataset, LTHNet achieved a 0.82%-2.31% improvement in performance and ACHNet achieved a 0.85%-2.17% improvement in performance. This is especially significant on unbalanced benchmarks such as IF = 100, where LTHNet and ACHNet achieve 1.95%-3.06%

Table 3. Retrieval performance of all methods compared on Cifar-100 for different IF and code length settings.

Imbalance Factor	IF1			IF50			IF100		
hash bits	32bits	64bits	96bits	32bits	64bits	96bits	32bits	64bits	96bits
DPSH	0.3113	0.4506	0.4957	0.1069	0.1407	0.1634	0.0978	0.1216	0.1383
HashNet	0.4380	0.5719	0.6311	0.1726	0.1950	0.2079	0.1444	0.1559	0.1631
DSDH	0.5398	0.6100	0.6407	0.1119	0.1000	0.0999	0.0940	0.0872	0.0807
CSQ	0.7711	0.7984	0.7821	0.2221	0.2745	0.2669	0.1716	0.1992	0.1658
LTHNet(k=0)	0.8195	0.8336	0.8400	0.2427	0.3028	0.3309	0.1752	0.2240	0.2415
LTHNet(k=3)	0.8268	0.8416	0.8490	0.2687	0.3354	0.3484	0.1819	0.2376	0.2620
LTHNet-Q	**0.8424**	**0.8528**	**0.8586**	**0.2902**	**0.3537**	**0.3629**	**0.2125**	**0.2629**	**0.2815**
ACHNet	0.8218	0.8299	0.8314	0.3075	0.3624	0.3708	0.2246	0.2770	0.2957
ACHNet-Q	**0.8357**	**0.8404**	**0.8406**	**0.3264**	**0.3792**	**0.3841**	**0.2513**	**0.2987**	**0.3126**

and 1.69%-2.67% improvements respectively, which illustrates the effectiveness of our proposed quantization objective on long-tailed distribution data. On the balanced benchmark, our method also improves.

Note that the performance improvement of our method is greater for shorter code lengths. The reason for this is that shorter hash codes contain less information, and more information loss occurs when converting images to these hash codes, which means greater quantization error. In addition, the proposed method improves the performance on cifar-100 greater than ImageNet-100. This is because the image sizes in the cifar dataset are small and the information richness is insufficient, leading to more quantization error.

Table 4. Retrieval performance of all methods compared on ImageNet-100 for different IF and code length settings.

Imbalance Factor	IF1			IF50			IF100		
hash bits	32bits	64bits	96bits	32bits	64bits	96bits	32bits	64bits	96bits
DPSH	0.4887	0.6055	0.6514	0.2186	0.3125	0.3791	0.1788	0.2832	0.3468
HashNet	0.4410	0.6006	0.6421	0.3465	0.4034	0.4240	0.3101	0.3770	0.3800
DSDH	0.6554	0.7015	0.7231	0.2568	0.2617	0.2744	0.1841	0.2134	0.2429
CSQ	0.8507	0.8733	0.8657	0.6629	0.7022	0.6823	0.5989	0.5620	0.5495
LTHNet(k=0)	0.7924	0.8267	0.8382	0.7369	0.7804	0.7920	0.6771	0.7350	0.7528
LTHNet(k=3)	0.8142	0.8453	0.8592	0.7612	0.8007	0.8157	0.7146	0.7665	0.7828
LTHNet-Q	**0.8286**	**0.8554**	**0.8674**	**0.7810**	**0.8182**	**0.8288**	**0.7377**	**0.7820**	**0.7966**
ACHNet	0.8592	0.8702	0.8779	0.8265	0.8427	0.8472	0.7965	0.8128	0.8163
ACHNet-Q	**0.8722**	**0.8806**	**0.8864**	**0.8456**	**0.8588**	**0.8596**	**0.8182**	**0.8285**	**0.8308**

To highlight the performance of our method on the tail classes, we divide all classes into four equal parts in order and evaluate MAP on the fourth part in which all classes are tail classes (i.e., the last 25 classes). We compare the results of our method with original long-tailed hashing methods in Table 5 and Table 6, indicating the outstanding performance of our method specifically on tail classes.

Table 5. On the IF=100 benchmark of Cifar-100, the performance comparison of different methods on the tail classes.

Methods	32bits	64bits	96bits
LTHNet	0.0565	0.1194	0.1243
LTHNet-Q	**0.1013**	**0.1405**	**0.1446**
ACHNet	0.0953	0.1569	0.1764
ACHNet-Q	**0.1387**	**0.1814**	**0.1903**

Table 6. On the IF=100 benchmark of ImageNet-100, the performance comparison of different methods on the tail classes.

Methods	32bits	64bits	96bits
LTHNet	0.4493	0.5402	0.5660
LTHNet-Q	**0.4717**	**0.5590**	**0.5796**
ACHNet	0.6543	0.6782	0.6910
ACHNet-Q	**0.6765**	**0.7052**	**0.7098**

4.4 Complexity Analysis

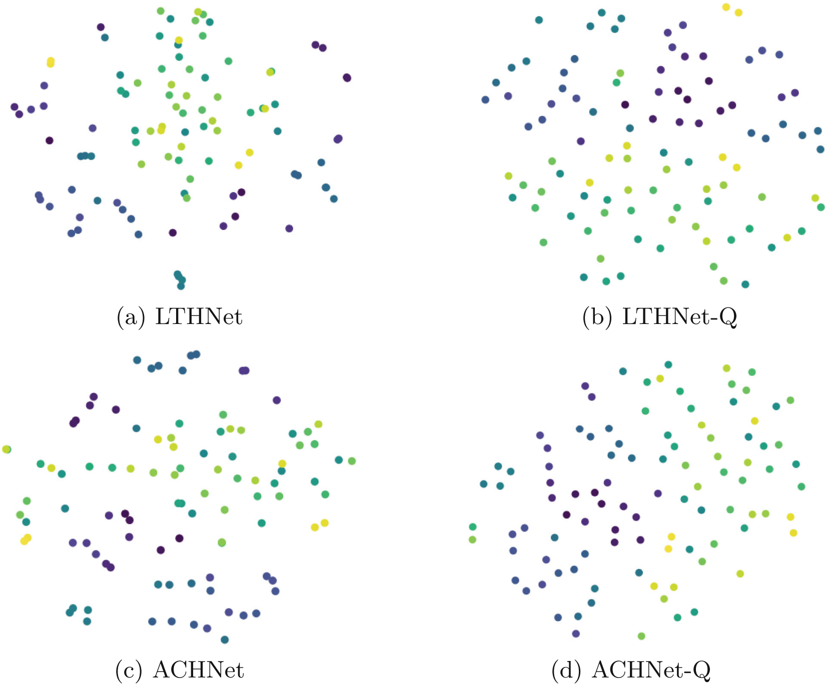
According to [12], the Sliced Wasserstein distance estimation in Eq. (6) has a computational complexity of $O(LN\log(Nd))$, where L is the number of random directions, N is the number of samples, and d is the dimension of the data. In most problems, Sliced Wasserstein distance requires a large number ($L \gg N$) of random directions, typically between 1000 to 10,000, to provide a reliable estimate of the distance [23]. To avoid random projections, we choose the direction that contains the information of the two distributions to be measured in Eq.(7), called non-random projection, which has a computational complexity of $O(mN\log(Nd))$. In this way, the number of directions is fixed to the dimension of the hashing space m, which is typically between 16 to 128 for many image hashing applications.

Taking ACHNet as an example, Table 7 reports the average running time for different quantization modes. We can observe that non-random projection is more computationally efficient than random projection, and relative running times have decreased by 20% to 30%. This is because it has fewer projections and omits the matrix-multiplication operation that projects the data points into random directions. In addition, there is a slight increase in the average running time compared to the original method, which is due to our additional calculation of the quantization loss.

Table 7. Average running time per epoch for different quantization modes (in seconds).

Datasets	Original	Random projection	Non-random projection
Cifar-100	19.5	27.3	21.7
ImageNet-100	51.4	68.1	55.6

4.5 Quantization Visualization


(a) LTHNet

(b) LTHNet-Q

(c) ACHNet

(d) ACHNet-Q

Fig. 4. Two-dimensional t-SNE visualization of hash centers.

In Fig. 4, we show the visualization of the hash centers of 100 classes in LTHNet and ACHNet, as well as after adding the quantization objective to them. For the generation of these hash centers, we simply take the average of all the hash codes of each class generated by the learned hash function as a prototype of this class. We project these hash centers into a two-dimensional space using t-SNE dimensionality reduction visualization. As shown in Fig. 4(a) and (c), the distribution of hash centers of the original method is very tight, and some hash centers even overlap, which usually belong to the tail classes. In contrast, the hash centers generated from the hash functions learned by the method after adding the quantization objective have better interclass separation, which alleviates the false negative problem described in 3.2 and improves the retrieval performance.

4.6 Sensitivity Analysis

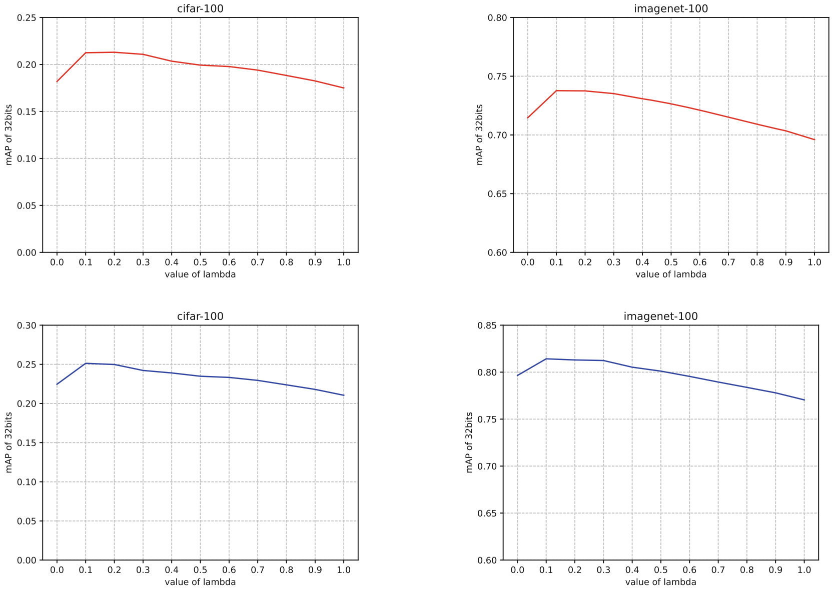


Fig. 5. (up) Sensitivity analysis of the parameter λ at 32bits on LTHNet-Q. (down) Sensitivity analysis of the parameter λ at 32bits on ACHNet-Q.

We conducted experiments on the sensitivity analysis of the parameter λ in Eq. (8). We choose LTHNet-Q and ACHNet-Q as the model and set the code length to 32bits. The experiment results are shown in Fig. 5. λ is 0 which is the original LTHNet and ACHNet method, and our proposed method achieves the best performance when λ is 0.1. Close performance is achieved when λ is 0.2 and 0.3. But when λ is larger than 0.3, the performance of the model gradually decreases, indicating that our method is more sensitive to λ. This is because when the value of λ is too large, the two models will be over-quantized, affecting their own performance. Therefore, we set the value of λ in Eq. (8) to 0.1.

5 Conclusion

In this paper, we propose a quantization objective for existing long-tailed hashing methods that do not consider the quantization error well. First, we take an optimal uniform discrete distribution as the target and minimize the distance between the continuous hash distribution and this target distribution. Then, we use the Sliced Wasserstein distance, which is easy to optimize and has low computational complexity, as a distance metric between the two distributions. Through comparative experiments, we verify the effectiveness of the proposed method, which significantly improves the performance of existing long-tailed hashing methods on two datasets.

Acknowledgements. This work was supported by Guangdong Basic and Applied Basic Research Foundation (2023A1515011400). The corresponding author is Yan Pan.

References

1. Arjovsky, M., Bottou, L.: Towards principled methods for training generative adversarial networks. In: International Conference on Learning Representations (2017)
2. Arjovsky, M., Chintala, S., Bottou, L.: Wasserstein generative adversarial networks. In: International conference on machine learning. pp. 214–223. PMLR (2017)
3. Cao, Z., Long, M., Wang, J., Yu, P.S.: Hashnet: Deep learning to hash by continuation. In: Proceedings of the IEEE international conference on computer vision. pp. 5608–5617 (2017)
4. Chawla, N.V., Bowyer, K.W., Hall, L.O., Kegelmeyer, W.P.: Smote: synthetic minority over-sampling technique. Journal of artificial intelligence research **16**, 321–357 (2002)
5. Chen, Y., Hou, Y., Leng, S., Zhang, Q., Lin, Z., Zhang, D.: Long-tail hashing. In: Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval. pp. 1328–1338 (2021)
6. Cui, Y., Jia, M., Lin, T.Y., Song, Y., Belongie, S.: Class-balanced loss based on effective number of samples. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 9268–9277 (2019)
7. Deshpande, I., Zhang, Z., Schwing, A.G.: Generative modeling using the sliced wasserstein distance. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3483–3491 (2018)
8. Doan, K.D., Manchanda, S., Mahapatra, S., Reddy, C.K.: Interpretable graph similarity computation via differentiable optimal alignment of node embeddings. In: Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval. pp. 665–674 (2021)
9. He, J., Chang, S.F., Radhakrishnan, R., Bauer, C.: Compact hashing with joint optimization of search accuracy and time. In: CVPR 2011. pp. 753–760. IEEE (2011)

10. Huang, C.Q., Yang, S.M., Pan, Y., Lai, H.J.: Object-location-aware hashing for multi-label image retrieval via automatic mask learning. IEEE Trans. Image Process. **27**(9), 4490–4502 (2018)
11. Kang, Q., Chen, X., Li, S., Zhou, M.: A noise-filtered under-sampling scheme for imbalanced classification. IEEE transactions on cybernetics **47**(12), 4263–4274 (2016)
12. Kolouri, S., Nadjahi, K., Simsekli, U., Badeau, R., Rohde, G.: Generalized sliced wasserstein distances. Advances in neural information processing systems **32** (2019)
13. Kou, X., Xu, C., Yang, X., Deng, C.: Attention-guided contrastive hashing for long-tailed image retrieval. In: IJCAI. pp. 1017–1023 (2022)
14. Krizhevsky, A., Hinton, G., et al.: Learning multiple layers of features from tiny images (2009)
15. Lai, H., Pan, Y., Liu, Y., Yan, S.: Simultaneous feature learning and hash coding with deep neural networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3270–3278 (2015)
16. Lai, H., Yan, P., Shu, X., Wei, Y., Yan, S.: Instance-aware hashing for multi-label image retrieval. IEEE Trans. Image Process. **25**(6), 2469–2479 (2016)
17. Li, Q., Sun, Z., He, R., Tan, T.: Deep supervised discrete hashing. Advances in neural information processing systems **30** (2017)
18. Li, W.J., Wang, S., Kang, W.C.: Feature learning based deep supervised hashing with pairwise labels. In: IJCAI. pp. 1711–1717 (2016)
19. Liang, Y., Pan, Y., Lai, H., Liu, W., Yin, J.: Deep listwise triplet hashing for fine-grained image retrieval. IEEE Trans. Image Process. **31**, 949–961 (2021)
20. Lin, T.Y., Goyal, P., Girshick, R., He, K., Dollár, P.: Focal loss for dense object detection. In: Proceedings of the IEEE international conference on computer vision. pp. 2980–2988 (2017)
21. Liu, W., Wang, J., Ji, R., Jiang, Y.G., Chang, S.F.: Supervised hashing with kernels. In: 2012 IEEE conference on computer vision and pattern recognition. pp. 2074–2081. IEEE (2012)
22. Liu, Z., Miao, Z., Zhan, X., Wang, J., Gong, B., Yu, S.X.: Large-scale long-tailed recognition in an open world. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 2537–2546 (2019)
23. Nguyen, K., Ho, N., Pham, T., Bui, H.: Distributional sliced-wasserstein and applications to generative modeling. arXiv preprint arXiv:2002.07367 (2020)
24. Shen, F., Shen, C., Liu, W., Tao Shen, H.: Supervised discrete hashing. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 37–45 (2015)
25. Wang, L., Pan, Y., Liu, C., Lai, H., Yin, J., Liu, Y.: Deep hashing with minimal-distance-separated hash centers. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 23455–23464 (2023)
26. Wang, Y.X., Ramanan, D., Hebert, M.: Learning to model the tail. Advances in neural information processing systems **30** (2017)
27. Weiss, Y., Torralba, A., Fergus, R.: Spectral hashing. Advances in neural information processing systems **21** (2008)
28. Xia, R., Pan, Y., Lai, H., Liu, C., Yan, S.: Supervised hashing for image retrieval via image representation learning. In: Proceedings of the AAAI conference on artificial intelligence. vol. 28 (2014)
29. Yuan, L., Wang, T., Zhang, X., Tay, F.E., Jie, Z., Liu, W., Feng, J.: Central similarity quantization for efficient image and video retrieval. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 3083–3092 (2020)

Unsupervised Metric Learning for Expressing Color and Shape Information to Uncover Abstract Connections within Image Datasets

Shun Obikane[1,2](✉), Haruna Tagawa[2], and Yoshimitsu Aoki[2]

[1] KOSÉ Corporation, 48-18 Sakae-cho, Kita-Ku, Tokyo 114-0005, Japan
shun_obikane@kose.co.jp

[2] Keio University, 3-14-1 Hiyoshi, Kohoku-ku, Yokohama, Kanagawa 223-8522, Japan
htagawa@aoki-medialab.jp, aoki@elec.keio.ac.jp

Abstract. In this research, we propose a novel approach using unsupervised metric learning tailored to datasets characterized by complex similarities and connections, such as those found in paintings and makeup, which are challenging to express linguistically. These datasets often present the difficulty of adequately analyzing data points due to the intricate interplay of defining elements, a limitation of traditional labeling methods. Additionally, the high degree of specialization required makes annotation significantly costly. Unsupervised metric learning emerges as a powerful method for extracting more cost-effective features and for the comprehensive analysis of these datasets. Expanding upon previous research that utilized style transfer models, our study further explores feature design, specifically focusing on extracting detailed information about critical aspects of similarity assessment, such as color and shape. Our model adeptly incorporates visual information, unveiling the hidden abstract connections within datasets. We validated our approach using a dataset of Ukiyo-e, a genre of Japanese painting, and achieved accuracy comparable to supervised learning models. This research opens up new possibilities for the analysis of complex image datasets with abstract relational depth, fostering a deeper understanding of the data.

Keywords: Unsupervised Metric Learning · Representation Learning · Image Retrieval Model

1 Introduction

In recent years, the significant progress of social media has led to the generation of vast amounts of image data. Effectively analyzing these data can offer numerous benefits for businesses and researchers. For instance, cluster analysis can identify trending patterns and help us understand people's interests and behavioral patterns. With advanced deep learning technology, analyzing these complex

A. Antonacopoulos et al. (Eds.): ICPR 2024, LNCS 15321, pp. 15–30, 2025.
https://doi.org/10.1007/978-3-031-78305-0_2

high-dimensional image datasets has become interpretable, allowing analysts to transform them into lower-dimensional, valuable features. Metric learning, one method that enables such feature extraction, involves learning the similarity between data points. However, defining a similarity that suits the analysis target is crucial. Metric learning primarily presupposes supervised learning. However, annotation costs can become quite significant when expert knowledge is required for similarity judgments or when the target dataset encompasses a wide range of classes. Therefore, unsupervised metric learning is more desirable. This research focuses on tasks where judging similarity is particularly challenging and aims to build a model using unsupervised metric learning to reduce the cost of feature generation while providing valuable information for image analysis. Unsupervised metric learning is advantageous for its adaptability to data requiring expert knowledge and being difficult to articulate. In the study of image similarity, themes such as painting and makeup are identified as examples that entail significant annotation costs. Paintings are similar in many elements, including the use of color and thematic choices, and are characterized by intricate similarities that arise from the subtle differences in techniques of individual artists. Similarly, makeup reflects individual creativity and fashion trends through colors, application techniques, and styles, and it is profoundly similar. In this paper, we define such similarity as "abstract connections". Utilizing unsupervised metric learning allows for the clever use of visual information in images without specifying similarity, revealing hidden abstract connections within the dataset through cluster analysis and promoting a deeper understanding of the data. Existing research on unsupervised metric learning primarily involves performing image transformations to create pairs based on the idea that the semantic meaning remains unchanged. However, this approach is unsuitable for tasks that require the detection of subtle differences in similarity. In these tasks, transformations could result in images with entirely different meanings. Previous studies explored image similarity centered on "color" and "shape" information [9,29]. Therefore, this research proposes a new unsupervised metric learning model that uses "color" and "shape" information to define abstract similarity effectively, representing these aspects efficiently.

2 Related Works

2.1 Supervised Metric Learning

Metric learning has been widely adopted as a feature extractor. Unlike classification models, which use the category of each image as supervised data, metric learning uses the similarity between images as the criterion. This approach directly targets how similar images are to each other. The main strength of metric learning is its focus on the relationships between data, which enables it to adapt effectively to scenarios such as class-imbalanced datasets and few-shot learning. This approach has been broadly applied in areas such as image retrieval model [4,16] and facial recognition [14]. Many methods have been proposed under the framework of supervised learning. Contrastive loss [8], triplet loss [25], and their advanced forms [23,28] are representative methods where the focus is placed on the design of input data based on labels of image similarity. Furthermore, by adding an L2 normalization layer immediately before the final layer of existing classification models [17], an approach has been introduced that allows metric learning to be advanced while maintaining classification learning. These advancements have led to attention being focused on the meticulous design of features for embedding data on hypersphere space, resulting in the development of methods [6,13,22,24] such as ArcFace [6]. While suited to tasks with clear classification categories, supervised metric learning encounters difficulties in fields like art and makeup, where the definition of similarity necessitates expert evaluations and incurs significant annotation costs. In such scenarios, unsupervised metric learning is considered a more favorable method.

2.2 Unsupervised Metric Learning

As the introduction mentions, fields such as art and makeup feature similarities based on diverse cultural backgrounds, suggesting hidden, complex relationships between data. Adopting unsupervised metric learning is essential to unravel these abstract connections. Unsupervised metric learning and representation learning have been proposed [2,3,12,27]. The majority of these methods are based on an approach where pseudo labels are employed for similarity labels. For instance, such methods have been developed that exploit the property where the semantic meaning of an image remains unchanged even when rotated [2], and techniques using algorithms such as keypoint matching to define identical images [12]. In the context of clustering tasks, as demonstrated by proposals such as Deep Cluster [3], utilizing the results of K-means clustering as pseudo labels allows for realizing unsupervised learning. This approach harnesses clustering outcomes as pseudo labels to facilitate the learning process. We focus on the similarity of fine details within images. Consequently, it becomes necessary to refrain from any manipulations of the original images. Furthermore, the challenge of dealing with data that embodies abstract relationships makes the presetting of cluster numbers impractical. Learning models that require the pre-establishment of cluster numbers are not suitable for addressing such problems due to inappropriate learning costs. Therefore, cluster analysis should be undertaken as a separate task following the generation of features. For these reasons, the necessity arises for conducting unsupervised metric learning without resorting to clustering-based pseudo labels or manipulated images.

2.3 Feature Extractor with Style/Makeup Transfer Model

From the perspective of feature extractors, early image retrieval models utilized image histograms to define similarity based on color and shape [9,29]. We adopted this approach, effectively extracting information on color and shape to evaluate their respective similarities. This concept is efficient for establishing criteria for similarity assessment in unsupervised situations. Moreover, in the context of similar image retrieval challenges, unsupervised learning methods using intermediate features of style transfer models have been proposed for datasets with abstract relationships like paintings [1,15]. We aimed to further develop this approach within the framework of unsupervised metric learning, focusing on extracting information on color and shape to create more practical models. To separately and appropriately handle color and shape, the use of the makeup transfer model [5,10,11,26], a form of style transfer, is considered adequate. Makeup transfer is the process of applying makeup from reference images to face images in source images. This model includes source and reference encoders, similar to style transfer, which are combined and processed through a decoder to produce the result image. Unlike the challenges in style transfer, this requires providing makeup colors from reference images while maintaining the face from the source image, demanding a specific output. Navigating the intricacies of facial features and adapting to a wide range of makeup styles present significant challenges in makeup. Thus, the model needs precise makeup color information from references and shape information from source images. This approach aligns closely with our study's concept, indicating a suitable direction for extending makeup transfer into unsupervised metric learning. Makeup transfer methods, built on adversarial learning, propose various approaches to accommodate diverse makeup styles. In reconsidering the roles of each encoder in makeup transfer, it is observed that the source encoder extracts shape information from source images, while the reference encoder extracts color information from reference images. We aim to further develop these encoders within the unsupervised metric learning framework to enhance them as effective feature extractors. This allows for the construction of models that can precisely identify similarities based on both shape and color. Through this approach, a broader exploration of similarity becomes feasible, opening new possibilities in similarity detection through unsupervised learning methods.

3 Method

3.1 Overview of our method

Let $\boldsymbol{X} = \{x_1, x_2, ...x_N\} \subset \mathbb{R}^{H\times W\times 3}$, where H is the height of the image, and W is the width, be the entire target data. Based on these data, a makeup transfer model is created. $\boldsymbol{I}_{src} \in \boldsymbol{X}$, which represents the source image to be changed, and a reference image $\boldsymbol{I}_{ref} \in \boldsymbol{X}$ with the change information (i.e., makeup information assuming that makeup is the task) are sampled from $\boldsymbol{X}$ in the same domain. The makeup transfer model $\boldsymbol{G}$ is created based on the paired data $\{\boldsymbol{I}_{src}, \boldsymbol{I}_{ref}\}$. Pairs of source images and reference images are randomly selected in the dataset. A characteristic of the model $\boldsymbol{G}$ is the inclusion of an L2 normalization layer between the encoder and decoder. As shown in Fig. 1, each encoder part of model $\boldsymbol{G}$ is used to perform the final learning as a feature generator. The proposed method comprises three learning steps. **Step 1** involves pretraining for makeup transfer to make generator $\boldsymbol{G}$ suitable for the face image task. Then, **Step 2** involves pretraining for metric learning to make the generator $\boldsymbol{G}$ suitable

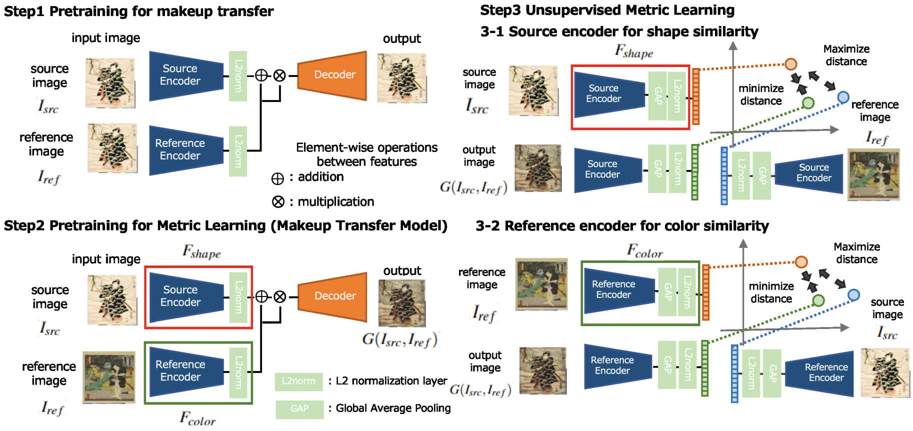

Step1 Pretraining for makeup transfer
input image
output
source image
I_{src}
reference image
I_{ref}
Source Encoder
Reference Encoder
L2norm
Decoder
Element-wise operations between features
⊕ : addition
⊗ : multiplication
Step2 Pretraining for Metric Learning (Makeup Transfer Model)
F_{shape}
F_{color}
$G(I_{src}, I_{ref})$
L2norm : L2 normalization layer
GAP : Global Average Pooling
Step3 Unsupervised Metric Learning
3-1 Source encoder for shape similarity
Maximize distance
minimize distance
output image
GAP
3-2 Reference encoder for color similarity

Fig. 1. Our method $\boldsymbol{G}$ is composed of three steps. Our model $\boldsymbol{G}$ is structured by connecting the source encoder, reference encoder, and decoder with an L2 normalization layer. In Step 1, the model solves an image reconstruction task to facilitate the learning process in Step 2. In Step 2, the model prepares for unsupervised metric learning in Step 3 by training within the makeup transfer framework. In Step 3, the model separates the source encoder$\boldsymbol{F}_{shape}$ and reference encoder$\boldsymbol{F}_{color}$, training them individually as feature generators to extract shape and color information, respectively. The encoders and the results $\boldsymbol{G}(\boldsymbol{I}_{src}, \boldsymbol{I}_{ref})$ generated in the makeup transfer model are effectively utilized to achieve unsupervised metric learning.

for makeup transfer. Since features are created using a makeup transfer model, we created a model that functions like this model does. Finally, in **Step 3**, the encoders that comprise the generator are extracted as feature extractor $\boldsymbol{F}_{shape}$ for shape information and feature extractor $\boldsymbol{F}_{color}$ for color information, and a metric is used for each of these feature extractors$\boldsymbol{F}_{shape}, \boldsymbol{F}_{color}$.

3.2 Step 1: Pretrain for makeup transfer model

First, to facilitate learning the makeup transfer task on step 2, we train generator $\boldsymbol{G}$ similar to an autoencoder model to pre-train a model suitable for the facial image task. The same image $\boldsymbol{I}$ is inputted to the source and reference encoders, and is trained with the aim of generating the same image as the input image, using the same image $\boldsymbol{I}$ as the output label. The loss in Step 1 is represented as follows:

$$L_{step1} = \|\boldsymbol{I} - \boldsymbol{G}(\boldsymbol{I}, \boldsymbol{I})\|_2. \tag{1}$$

3.3 Step 2: Pretrain for metric learning (makeup transfer model)

Through the learning process in Step 1, generator $\boldsymbol{G}$ becomes a model specialized for the topic of the targeted dataset (for example, if it's makeup dataset, the task is face images, and if it's paintings, the task is artwork images). We focused on the characteristics of the makeup transfer model; the makeup transfer task is solved in Step 2 based on the results in Step 1. The loss function was designed to effectively achieve makeup transfer, which involves providing the color information of the reference image $\boldsymbol{I}_{ref}$ to the target part of the source image $\boldsymbol{I}_{src}$; its design is detailed below.

Adversarial loss : Adversarial loss is used to create natural output results. Following the general adversarial loss approach, we introduce a discriminator $\boldsymbol{D}$ to distinguish whether the input image is a generated image $\boldsymbol{G}(\boldsymbol{I}_{src}, \boldsymbol{I}_{ref})$ or a source image$\boldsymbol{I}_{src}$, and train the generator $\boldsymbol{G}$ to output realistic images that can deceive the discriminator $\boldsymbol{D}$. The entire adversarial loss can be described as follows:

$$L^{D}_{adv} = \mathbb{E}[\log \boldsymbol{D}(\boldsymbol{I}_{src})] + \mathbb{E}[\log(1 - \boldsymbol{D}(\boldsymbol{G}(\boldsymbol{I}_{src}, \boldsymbol{I}_{ref})))], \quad (2)$$

$$L^{G}_{adv} = -\mathbb{E}[\log(\boldsymbol{D}(\boldsymbol{G}(\boldsymbol{I}_{src}, \boldsymbol{I}_{ref}))]. \quad (3)$$

Makeup loss : In makeup transfer, images that reflect the results to a certain extent are used as the pseudo-ground truth. Histogram Matching (HM) [7], which reflects the makeup of a reference image with respect to a source image, is often used for makeup transfer. $HM(x, y)$ is a method that creates an image with the same color distribution as that of images x and y, while preserving the identity of x. It is named "makeup loss" but the same process will be applied to all datasets.

$$L_{makeup} = \|HM(\boldsymbol{I}_{src}, \boldsymbol{I}_{ref}) - \boldsymbol{G}(\boldsymbol{I}_{src}, \boldsymbol{I}_{ref})\|_2. \quad (4)$$

Perceptual loss : The identity of the source image should be preserved even after transfer. The perceptual loss function is typically used for style transfers. In our method, the source encoder $\boldsymbol{F}_{shape}$ side of the model created in Step 1 is used, some features are extracted in the same manner, and the perceptual loss is calculated. The shape encoder l-th layer is defined as $\boldsymbol{F}_{shape-l}$, and L layers are used; the perceptual loss is described as follows:

$$L_{per} = \sum_{l}^{L} \|\boldsymbol{F}_{shape-l}(\boldsymbol{G}(\boldsymbol{I}_{src}, \boldsymbol{I}_{ref})) - \boldsymbol{F}_{shape-l}(\boldsymbol{I}_{src})\|_2. \quad (5)$$

Feature matching loss : In style transfer, which is the basis of makeup transfer, the matching of the output feature values is set as a loss function (style loss, content loss). Our method defines the feature matching losses L^{shape}_{feat}, L^{color}_{feat} from the viewpoints of shape and color matching, respectively by the cosine similarity $cos(x, y) = \frac{x^T, y}{\|x\|\|y\|}$. Regarding shape L^{shape}_{feat}, the source image $\boldsymbol{I}_{src}$ and generated image $\boldsymbol{G}(\boldsymbol{I}_{src}, \boldsymbol{I}_{ref})$ should have the same shape; accordingly, the following relation is obtained:

$$L^{shape}_{feat} = 1 - cos(\boldsymbol{F}_{shape}(\boldsymbol{G}(\boldsymbol{I}_{src}, \boldsymbol{I}_{ref})), \boldsymbol{F}_{shpae}(\boldsymbol{I}_{src})). \quad (6)$$

On the other hand, regarding color $L_{feat-color}$, the reference image $\boldsymbol{I}_{ref}$ and generated image $\boldsymbol{G}(\boldsymbol{I}_{src}, \boldsymbol{I}_{ref})$ should have the same color; therefore, the following relation is obtained:

$$L_{feat}^{color} = 1 - cos(\boldsymbol{F}_{color}(\boldsymbol{G}(\boldsymbol{I}_{src}, \boldsymbol{I}_{ref})), \boldsymbol{F}_{color}(\boldsymbol{I}_{ref})). \quad (7)$$

Our loss functions for $\boldsymbol{G}$ and $\boldsymbol{D}$ are represented as follows:

$$L_G = \lambda_{adv} L_{adv}^G + \lambda_{per} L_{per} + \lambda_{makeup} L_{makeup} + \lambda_{feat}^{shape} L_{feat}^{shape} + \lambda_{feat}^{color} L_{feat}^{color} \quad (8)$$

$$L_D = \lambda_{adv} L_{adv}^D \quad (9)$$

where $\lambda_{adv}, \lambda_{per}, \lambda_{makeup}, \lambda_{feat}^{shape}, \lambda_{feat}^{color}$ are hyperparameters.

3.4 Step 3: Unsupervised Metric Learning

Following previous studies, it has been shown that features, including those from style transfer (makeup transfer) models created at Step 2, are effective for measuring similarity. At Step 3, we further refine the model obtained from Step 2 as a feature extractor. Our method employs unsupervised metric learning with pseudo labels related to similarity. Such meticulous adjustments allow for the creation of models that are adaptable even with a small number of samples. Initially, a single model $\boldsymbol{G}$ is divided into two models: $\boldsymbol{F}_{shape}$, for extracting shape information, and $\boldsymbol{F}_{color}$, for extracting color information, with each undergoing further training. For the sake of convenience in notation, the output result $\boldsymbol{G}(\boldsymbol{I}_{src}, \boldsymbol{I}_{ref})$ will be denoted as $\boldsymbol{I}_{output}$. The fundamental idea of our method, similar to the feature matching loss at Step 2, focuses on the similarity of color and shape information between the generated output and input data, creating pairs that can be identified as similar or dissimilar. For $\boldsymbol{F}_{shape}$, since the input data $\boldsymbol{I}_{src}$ and the generated result $\boldsymbol{I}_{output}$ represent the same subject, they are treated as identical pairs in terms of shape information. This is defined as Eq. (10). Since $\boldsymbol{I}_{src}$ and $\boldsymbol{I}_{ref}$ could potentially form a pair with low similarity, they are treated as dissimilar pairs. This is defined as Eq. (11).

$$L_{sim}^{shape} = 1 - cos(\boldsymbol{F}_{shape}(\boldsymbol{I}_{src}), \boldsymbol{F}_{shape}(\boldsymbol{I}_{output})) \quad (10)$$

$$L_{dissim}^{shape} = 1 + cos(\boldsymbol{F}_{shape}(\boldsymbol{I}_{src}), \boldsymbol{F}_{shape}(\boldsymbol{I}_{ref})) \quad (11)$$

In $\boldsymbol{F}_{color}$, the input data $\boldsymbol{I}_{ref}$ and the generated result $\boldsymbol{I}_{output}$, sharing the same coloration, are treated as identical in terms of color information. This is also defined as Eq. (12). Similarly, $\boldsymbol{I}_{ref}$ and $\boldsymbol{I}_{src}$ are treated as dissimilar pairs, as discussed for $\boldsymbol{F}_{shape}$. This is defined as Eq. (13).

$$L_{sim}^{color} = 1 - cos(\boldsymbol{F}_{color}((\boldsymbol{I}_{ref}), \boldsymbol{F}_{color}(\boldsymbol{I}_{output})) \quad (12)$$

$$L_{dissim}^{color} = 1 + cos(\boldsymbol{F}_{color}(\boldsymbol{I}_{ref}), \boldsymbol{F}_{color}(\boldsymbol{I}_{src})) \quad (13)$$

The loss function designed to separate dissimilar pairs ceases to apply after a certain number of epochs, $epoch_{stop}$. This is because $\boldsymbol{I}_{src}$ and $\boldsymbol{I}_{ref}$, defined as dissimilar pairs, might potentially form a similar pair, and ultimately, the focus shifts solely to learning from identical pairs to enhance the accuracy of similarity definitions.

4 Experiments

4.1 Dataset

We utilized the ARC Ukiyo-e Faces Dataset [19], representing one of the genres of Japanese paintings known as Ukiyo-e. As labels, we adopted "painters" and focused on the abstract similarities in their unique painting techniques. The data possessing abstract connections, which is the focus of our study, frequently appears in various applications. In general image datasets, when performing clustering analysis, the critical factor is often shape information alone. For example, even if the color is different, the same car would be assigned to the "car" class. However, in datasets with abstract relationships, color information also contributes to similarity assessment and becomes a similarity criterion. It is necessary to predict the background and measure similarity based on both color and shape information. In cases where the dataset does not have a general categorization and belongs to an unknown and difficult-to-explain category, it may exhibit similarities that differ from general datasets, as it is unclear whether color or shape information is the key factor. Therefore, compared to standard datasets, a more abstract and complex form of similarity is required. However, there are limited datasets available for quantitative evaluation. We identified that evaluation using data labeled with "painters" is suitable for the concept of data with abstract connections. Although validation with more general data is also necessary, our study specifically targets data with abstract and complex connections. Therefore, the evaluation was conducted using this dataset. Due to insufficient samples for proper evaluation in some categories, we selected 17 categories, each containing over 80 images, as labels. Fifty images were randomly chosen from each class as test data. The final dataset consisted of 3,195 images for training and 850 images for testing.

4.2 Implementation detail

All the learning steps were trained with the Adam optimizer (β_1=0.500, β_2=0.999). The learning rate for the generator in Step 1 was 5.0×10^{-2}, that for both the generator and discriminator in Step 2 were 1.0×10^{-3}, and that in Step 3 was 5.0×10^{-4}. All the batch sizes were 32. The number of epochs in Step 1, Step 2, and Step 3 was 200, 100, and 100, respectively. The encoder utilizes a model similar to VGG [18], while the decoder is a five-layer model employing deconvolution layers. In both the encoder and decoder, instance normalization [20] is used as the batch normalization layer. In our method, the dimensionality of the feature vectors is set to 512 dimensions. The network was implemented using PyTorch with a single NVIDIA RTX A5000 GPU. Regarding the hyperparameters, they were $\lambda_{adv} = 1, \lambda_{per} = 1, \lambda_{makeup} = 100, \lambda_{feat}^{shape} = 1, \lambda_{feat}^{color} = 1, \alpha = 10, epoch_{stop} = 50$.

5 Result

5.1 Evaluation method

Performance assessments were conducted to evaluate the feature representation of our method, focusing on its application in recommendation models and clustering analysis. For the recommendation model, Precision@K (where K represents the number of reference images) was utilized as a measure of accuracy. During training, no labels were used; however, for evaluation, labels were assigned to the training data as reference images as a recommendation model. The most similar training data to each test data point was determined using cosine similarity, and the class related to the painter was estimated. This approach allowed us to verify whether the proposed method captured information relevant to similarity. In the case of clustering analysis, Spectral Clustering [21] was applied to the generated feature sets, and Normalized Mutual Information (NMI) was used as the metric for evaluation. As for the accuracy comparison, our method is an unsupervised learning technique, and the accuracy in the presence of actual supervised data was set as the target. ArcFace [6], a standard method in supervised metric learning, was used as a comparison. Additionally, well-known unsupervised learning methods such as DeepCluster [3], UEL [27], and UDMLSSS [2] were employed for comparative analysis. Each backbone model employed in this study utilizes ResNet50. The part that can be treated as an encoder is connected to a decoder, forming an autoencoder, and the image reconstruction task on the target dataset [19] is solved to perform pretraining. Through these evaluations, we assess whether the features generated by our method can aggregate useful information when applied to various analytical techniques. Compared to tasks with clearly defined general categories, the challenge of inferring the artist from paintings requires a nuanced judgment criterion and represents a highly specialized and challenging task setting. In experiments conducted with other methods, the subjects are often readily identifiable items such as products or biological categories, which differ in purpose. However, our evaluation design aims to demonstrate the effectiveness of our method in uncovering abstract connections within the targeted data by comparing it with previous methods. Our method considers two feature types: shape and color, which are integrated using a parameter μ $(0 \leq \mu \leq 1)$:

$$\mu \boldsymbol{F}_{shape}(\boldsymbol{I}) + (1 - \mu)\boldsymbol{F}_{color}(\boldsymbol{I}). \tag{14}$$

This parameter dictates the emphasis on shape or color, with a higher μ prioritizing shape information and a lower μ favoring color information.

5.2 Quantitative evaluation

Our method, unique in its reliance on unsupervised learning, is evaluated against a hypothetical scenario where training data are accompanied by supervised labels, a setting that represents ideal performance. For the benchmark in supervised learning, we employed the ArcFace [6] method. Our research follows a

Table 1. The results of the accuracy evaluation using Precision@K, tailored for the recommendation model, are presented. We used K values of 1, 5, 10, 20, and 30, with the maximum value of 30 sets considering the minimum sample size per class in the training data. The scores of each comparative method and the score at each learning step are shown. Moreover, in step 3, a loss function designed to distance the features is implemented. We provide results for our method without this loss function (our method (1)) and with it (our method (2)). While achieving numbers close to the supervised learning model ArcFace [6], our method also surpasses previous methods [2,3,27] in terms of score. Additionally, accuracy improvements at each learning stage indicate that each step effectively contributes to the overall performance. The increasing effectiveness of our method with larger values of K suggests the richness of the feature information and the excellent cohesion of the features.

No.	Model Type	Precision@1	Precision@5	Precision@10	Precision@20	Precision@30
1	ArcFace [6]	0.562	0.549	0.507	0.482	0.441
2	Deep Cluster [3]	0.315	0.315	0.287	0.271	0.250
3	UEL [27]	0.518	0.492	0.465	0.417	0.394
4	UDMLSS [2]	0.565	0.52	0.497	0.465	0.440
5	step 1 model	0.537	0.482	0.452	0.418	0.377
6	step 2 model	0.611	**0.572**	0.524	0.467	0.432
7	our method(1)	0.612	0.550	0.528	0.480	0.443
8	our method(2)	**0.616**	0.570	**0.550**	**0.497**	**0.472**

Table 2. Like Table 1, the results presented here use Precision@K for accuracy evaluation, designed with the recommendation model. This result demonstrates the accuracy within minor classes with limited sample sizes in the training data, precisely two classes with only 38 and 31 samples. These sample sizes represent approximately 1% of the total data and pose a significant challenge for unsupervised learning methods. The results indicate that our method is highly effective for these minor classes, a notable strength of our approach. Moreover, there are significant improvements across the learning stages (comparing step 1 and step 2), with a substantial impact of unsupervised metric learning demonstrated in step 3.

No.	Model Type	Precision@1	Precision@5	Precision@10	Precision@20	Precision@30
1	ArcFace [6]	0.340	0.240	0.150	0.090	0.040
2	Deep Cluster [3]	0.060	0.080	0.040	0.040	0.010
3	UEL [27]	0.300	0.270	0.200	0.140	0.080
4	UDMLSS [2]	0.420	0.290	0.290	0.230	0.170
5	step 1 model	0.410	0.320	0.210	0.130	0.100
6	step 2 model	0.450	0.400	0.350	0.190	0.150
7	our method(1)	0.490	0.410	0.400	0.280	0.210
8	our method(2)	**0.500**	**0.510**	**0.460**	**0.420**	**0.410**

systematic approach, involving three stages of learning steps. To verify the effectiveness of each step, we present the results as step 1, step 2, and our method, corresponding to each stage of learning. This allows for a clear comparison and

Table 3. The results of evaluating the functionality of feature representations in cluster analysis. The evaluation utilized the Normalized Mutual Information (NMI) as the scoring metric. Our method has achieved score close to that of ArcFace [6], a supervised learning technique, and has demonstrated good accuracy compared to other previous methods [2,3,27]. Furthermore, comparing the score between step 1 and step 2 shows a progressive improvement in scores, indicating that the features retain similarity information at each stage of learning while forming clusters effectively.

ArcFace [6]	Deep Cluster [3]	UEL [27]	UDMLSS [2]
0.4194	0.1558	0.3631	0.3885

step 1 model	step 2 model	our method
0.3319	0.3840	**0.3976**

Table 4. The results using various values of the parameter μ (0, 0.25, 0.50, 0.75, 1.00) are presented concerning the accuracy at Precision@1, 5, and 30. The accuracy varies depending on the value of μ. This result indicates that μ is a critical parameter that should be chosen based on whether shape or color information is more important for the task.

μ	0.00	0.25	0.50	0.75	1.00
Precision@1	0.320	0.381	0.507	0.616	**0.625**
Precision@5	0.296	0.329	0.562	**0.571**	0.565
Precision@30	0.236	0.413	0.468	**0.473**	0.461

evaluation of our approach. Step 1 functions under the same conditions as a conventional auto-encoder model; hence, the results are presented as such. Step 2 resembles the scenario of a makeup (style) transfer model and uses a similar setup to existing methods [1,15] that employ style transfer features. Our method represents the final proposed approach after completing all the learning steps. In step 3, in addition to the loss function that brings the pseudo-similar pairs of each source encoder and reference encoder closer, a loss function is also introduced to treat the source and reference images as temporary dissimilar pairs To evaluate the impact of this loss function, results without this loss function are noted as our method(1) and with the loss function as our method(2). The results of using Precision@K (where K is 1, 5, 10, 20, 30) for evaluating a recommendation model are presented in Table 1. Table 1 displays the results for all 17 classes considered. As shown in Table 1, the proposed method achieves scores comparable to ArcFace [6], a supervised learning method, indicating that it has reached a satisfactory score as an unsupervised learning method. Furthermore, the proposed method's scores are superior to those of previous methods [2,3,27]. Additionally, an examination of the results at each learning stage within the proposed method reveals improvement scores at each learning step, demonstrating the effectiveness of each training phase. Notably, as the number of reference images K increases, the superiority of the scores also rises, indicating that our method effectively

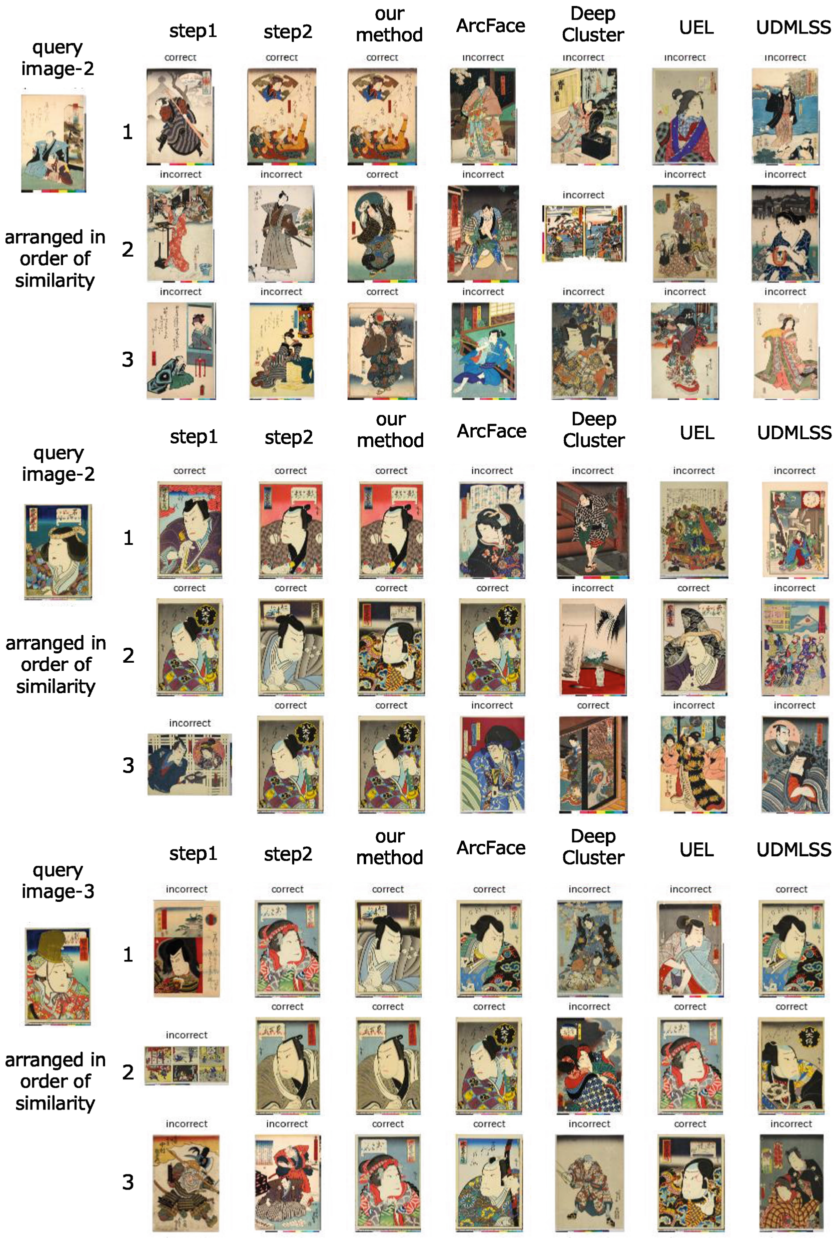


Fig. 2. Example of qualitative evaluation: The results of each method are listed in numerical order based on similarity to the query image. If the query image matches the painter, it is labeled as 'correct'; if it does not match, it is labeled as 'incorrect'.

groups images with similar information, whether at a smaller or larger scale. The strong results achieved in these unsupervised metric learning experiments are likely due to the characteristics of the targeted task. Table 2 shows the score for minor classes with limited sample sizes. Specifically, it targets two classes in the training data with only 38 and 31 samples, respectively. These sample sizes account for approximately 1% of the total data, making them particularly challenging for unsupervised learning. The model type listed in the results is the same as that in Table 2. Initially, compared to ArcFace, a supervised learning model, our method achieves comparable accuracy and, as the number of reference images K increases, significantly surpasses the ArcFace scores. This is consistent even compared to existing methods, demonstrating that our approach is highly effective for minor classes. The strength of our method in handling class imbalance issues stands out as one of its most significant advantages in unsupervised metric learning. Furthermore, substantial improvements in scores are observed at each learning stage, indicating that the method is particularly effective for minor classes. Our approach, which is robust against minor classes, enables the design of feature vectors that capture intricate information, proving highly effective in unraveling abstract connections within the dataset. Although providing supervised labels offers a more advantageous scenario for models, for the complex, highly specialized tasks involving intricate image elements and expected abstract connections targeted in this study, representing similarities based solely on detailed image information without labels may be more appropriate. The superiority of our method over previous methods lies in the deeper involvement of color information in the feature elements, as explained earlier regarding the dataset (Sec. 4.1). Compared to the general datasets targeted by previous methods, it is evident that carefully extracting and utilizing both shape and color information as features is effective in datasets with unique similarities. Table 3 presents the results using Normalized Mutual Information (NMI) as the metric for evaluating the application of the generated features in cluster analysis. As the results indicate, similar to the recommendation model-focused evaluations, our method achieves scores comparable to supervised scenarios and produces better clusters than existing methods. Our method can generate highly effective features that efficiently group similar entities. Table 4 presents the results of evaluating the parameter μ used to aggregate features. The value of μ was varied at 0, 0.25, 0.5, 0.75, and 1 to assess its impact quantitatively. This parameter μ indicates the emphasis on color information versus shape information: higher μ values prioritize shape information, while lower values favor color information. According to the results in Table 4, the highest accuracy was achieved when color and shape information were equally considered (μ=0.75), and there was also a relatively high accuracy trend when prioritizing shape information. Mainly for this dataset, the results suggest that shape information is a crucial determinant in similarity judgments. The importance of color versus shape may vary depending on the target dataset, necessitating adjustments to these parameters based on which factor is more critical for assessing similarities. We also investigated the aggregation parameter μ, varied from 0 to 1, to assess its quantitative impact. μ

indicates the emphasis on color or shape information: a higher μ leans towards shape, while a lower μ favors color. Results in Table 4 show the highest accuracy at an equal treatment of color and shape information ($\mu = 0.5$). It is indicated that color information plays a key role in similarity assessment for this dataset. Therefore, depending on the dataset, the importance of color versus shape must be adjusted for optimal similarity determination.

5.3 Qualitative evaluation

Fig. 2 presents actual examples for qualitative evaluation. For each query image, we displayed images with high similarity in order, based on cosine similarity. The results of our method at each learning stage (Steps 1, 2, and 3) were compared with those of the supervised learning model ArcFace [6] and previous unsupervised models [2,3,27], with Step 3 representing the final outcome of our approach. We observed a trend of features of images with high similarity progressively clustering at each learning stage. Notably, as with step 2 model like style transfer (makeup transfer) models, satisfactory results could sometimes be achieved as early as Step 2, as seen with Query Image 2. For quantitative evaluation, despite treating the artist as a label, our method was able to effectively extract images with similar styles, even when the artists differed a characteristic not present in ArcFace [6], a supervised learning model and previous unsupervised models [2,3,27]. This result suggests that our unsupervised metric learning-based method is highly useful for generating features to explore latent similarities and hidden connections within datasets. This study confirms the effectiveness of unsupervised metric learning in aggregating features of images with linguistically challenging and complex similarities and in uncovering hidden relationships within datasets. Future research is expected to expand the utility of this method through its application to a more diverse range of datasets.

6 Conclusion

In this research, we proposed an unsupervised metric learning model. This model is particularly effective in capturing similarities that are difficult to express in language and uncovering hidden connections within datasets, such as those involving paintings and makeup. We conducted experiments using a dataset of Ukiyo-e, a traditional form of Japanese art. By evaluating accuracy using the artists as labels, we achieved a precision close to that of supervised models, thereby proving the effectiveness of our approach. Moreover, our method involved a three-stage learning process, and we were able to confirm the effectiveness at each stage. Furthermore, the greatest strength of our method lies in its ability to handle minor classes effectively while keeping annotation costs low. Moving forward, we aim to explore deeper into the latent similarities and unravel hidden connections in a more diverse range of datasets.

References

1. Benitez-Garcia, G., Shimoda, W., Yanai, K.: Style image retrieval for improving material translation using neural style transfer. In: Proceedings of the 2020 Joint Workshop on Multimedia Artworks Analysis and Attractiveness Computing in Multimedia. pp. 1–6 (2020). https://doi.org/10.1145/3379173.3393707
2. Cao, X., Chen, B.C., Lim, S.N.: Unsupervised deep metric learning via auxiliary rotation loss. arXiv preprint arXiv:1911.07072 (2019)
3. Caron, M., Bojanowski, P., Joulin, A., Douze, M.: Deep clustering for unsupervised learning of visual features. In: European Conference on Computer Vision (2018)
4. Chen, W., Liu, Y., Wang, W., Bakker, E.M., Georgiou, T., Fieguth, P., Liu, L., Lew, M.S.: Deep learning for instance retrieval: A survey. IEEE Transactions on Pattern Analysis; Machine Intelligence **45**(06), 7270–7292 (2023)
5. Deng, H., Han, C., Cai, H., Han, G., He, S.: Spatially-invariant style-codes controlled makeup transfer. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 6549–6557 (June 2021)
6. Deng, J., Guo, J., Xue, N., Zafeiriou, S.: Arcface: Additive angular margin loss for deep face recognition. In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 4685–4694 (2019)
7. G., R.C., W., R.E.: Digital Image Processing. Pearson/Prentice Hall, 4th edn. (2017)
8. Hadsell, R., Chopra, S., LeCun, Y.: Dimensionality reduction by learning an invariant mapping. In: 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06). vol. 2, pp. 1735–1742 (2006)
9. Jain, A.K., Vailaya, A.: Image retrieval using color and shape. In: Pattern recognition, 29(8). pp. 1233–1244 (1996)
10. Jiang, W., Liu, S., Gao, C., Cao, J., He, R., Feng, J., Yan, S.: Psgan: Pose and expression robust spatial-aware gan for customizable makeup transfer. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (June 2020)
11. Li, T., Qian, R., Dong, C., Liu, S., Yan, Q., Zhu, W., Lin, L.: Beautygan: Instance-level facial makeup transfer with deep generative adversarial network. In: Proceedings of the 26th ACM international conference on Multimedia. pp. 645–653 (October 2018)
12. Li, Y., Kan, S., He, Z.: Unsupervised deep metric learning with transformed attention consistency and contrastive clustering loss. In: European Conference on Computer Vision (2020)
13. Liu, W., Wen, Y., Yu, Z., Li, M., Raj, B., Song, L.: Sphereface: Deep hypersphere embedding for face recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 212–220 (2017)
14. Masi, I., Wu, Y., Hassner, T., Natarajan, P.: Deep face recognition: A survey. In: 2018 31st SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI). pp. 471–478 (2018)
15. Matsuo, S., Yanai, K.: Cnn-based style vector for style image retrieval. In Proceedings of the 2016 ACM on International Conference on Multimedia Retrieval pp. 309–312 (2016)
16. Park, S., Shin, M., Ham, S., Choe, S., Kang, Y.: Study on fashion image retrieval methods for efficient fashion visual search. In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). pp. 316–319 (2019). https://doi.org/10.1109/CVPRW.2019.00042

17. Ranjan, R., Castillo, C.D., Chellappa, R.: L2-constrained softmax loss for discriminative face verification. arXiv preprint arXiv:1703.09507 (2017)
18. Simonyan, K., Z., A.: Very deep convolutional networks for large-scale image recognition. arXiv arXiv:1409.1556 (2014)
19. Tian, Y., Clanuwat, T., Suzuki, C., Kitamoto, A.: Ukiyo-e analysis and creativity with attribute and geometry annotation. In: Proceedings of the International Conference on Computational Creativity (2021)
20. Ulyanov, D., V., A., L., V.: Instance normalization: The missing ingredient for fast stylization. arXiv arXiv:1607.08022 (2016)
21. Von Luxburg, U.: A tutorial on spectral clustering. Stat. Comput. **17**, 395–416 (2007)
22. Wang, F., Cheng, J., Liu, W., Liu, H.: Additive margin softmax for face verification. IEEE Signal Process. Lett. **25**(7), 926–930 (2018)
23. Wang, F., Zuo, W., Lin, L., Zhang, D., Zhang, L.: Joint learning of single-image and cross-image representations for person re-identification. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1288–1296 (2016)
24. Wang, H., Wang, Y., Zhou, Z., Ji, X., Gong, D., Zhou, J., Liu, W.: Cosface: Large margin cosine loss for deep face recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 5265–5274 (2018)
25. Wang, J., song, Y., Leung, T., Rosenberg, C., Wang, J., Philbin, J., Chen, B., Wu, Y.: Learning fine-grained image similarity with deep ranking. Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (04 2014)
26. Xiang, J., Chen, J., Liu, W., Hou, X., Shen, L.: Ramgan: Region attentive morphing gan for region-level makeup transfer. In: In Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXII. pp. 719–735 (2022)
27. Ye, M., Zhang, X., Yuen, P.C., Chang, S.F.: Unsupervised embedding learning via invariant and spreading instance feature. In: IEEE International Conference on Computer Vision and Pattern Recognition (CVPR) (2019)
28. Yu, B., Liu, T., Gong, M., Ding, C., Tao, D.: Correcting the triplet selection bias for triplet loss. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 71–87 (2018)
29. Yue, J., Li, Z., Liu, L., Fu, Z.: Content-based image retrieval using color and texture fused features. In: Mathematical and Computer Modelling. vol. 54, pp. 1121–1127 (2011)

Fashion Image Retrieval with Occlusion

Jimin Sohn[1(✉)], Haeji Jung[2], Zhiwen Yan[3], Vibha Masti[3], Xiang Li[3],
and Bhiksha Raj[3]

[1] GIST, Gwangju, South Korea
estelle26598@gm.gist.ac.kr
[2] Korea University, Seoul, South Korea
gpwl0709@korea.ac.kr
[3] Carnegie Mellon University, Pittsburgh, USA
{vmasti,xl6,zhiweny}@andrew.cmu.edu, bhiksha@cs.cmu.edu

Abstract. With the growth of online fashion platforms and independent content creators, there is a growing interest in visually searching for similar clothing items as shown online. In real-world settings, clothes are often covered by other objects, making retrieval challenging. To make fashion image retrieval more robust, we explore fashion image retrieval with occlusion. We conducted various experiments on the In-shop Clothes Retrieval dataset, a subset of the DeepFashion benchmark. We constructed variations of the dataset with different occlusion types, including various sizes and locations of MSCOCO objects and object masks to simulate realistic occlusion circumstances. We evaluate the zero-shot and fine-tuned performance of the state-of-the-art models on these datasets and observe performance drop. We observe that fine-tuning models on one occluded dataset makes the model more robust to other occlusion types and reduces performance drop. The dataset used in this paper can be found in https://bit.ly/4749Mbo.

Keywords: Image Retrieval · Occlusion · Robust Model

1 Introduction

With the increasing influence of the fashion industry and independent content creators, Fashion Image Retrieval (FIR) [7,22,37,41] has emerged as a crucial task combining the fields of computer vision and fashion technology. FIR tasks differ from general Image Retrieval tasks specifically in practical settings, where input images may be corrupted or occluded. This becomes particularly relevant

J. Sohn, H. Jung, Z. Yan, and V. Masti—These authors contributed equally to this work.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/978-3-031-78305-0_3.

A. Antonacopoulos et al. (Eds.): ICPR 2024, LNCS 15321, pp. 31–46, 2025.
https://doi.org/10.1007/978-3-031-78305-0_3

when considering images captured by consumers, as they may encompass diverse viewpoints, shapes, and obstructions by unrelated objects.

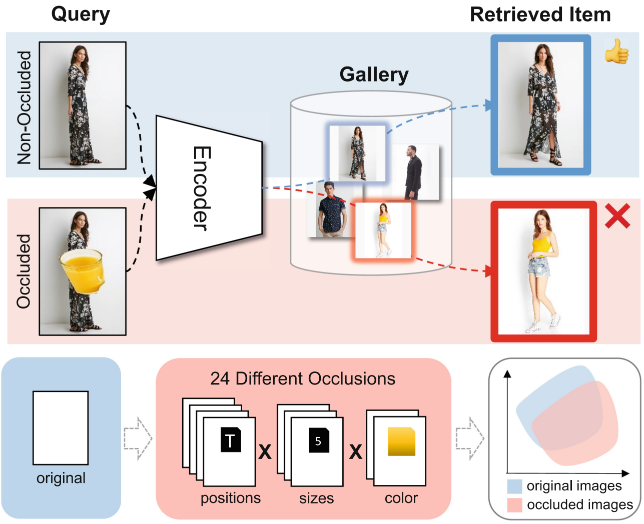


Fig. 1. Existing fashion image retrieval methods often fail to perform properly when given an occluded image. We investigate the impact of different properties of occlusions – positions, sizes, and colored types, by creating 24 different occluded datasets and analyzing the quantitative/qualitative results.

As illustrated in Fig.1, we first investigated the performance degradation of existing FIR models when confronted with occluded images as queries, using the models proposed by An [2] and Ermolov [11]. To conduct a more comprehensive analysis, we have constructed **IS-Occ(InShop-Occluded) Datasets** that incorporates different occluded versions of an existing benchmark dataset, InShop [24]. IS-Occ has two variations of the occluded InShop dataset: one with black object masks sourced from the MSCOCO [21] dataset, and the other with real objects also obtained from MSCOCO. Both versions showcase different types of occlusions, including variations in occlusion ratio and positioning within the image. These occlusions are designed to simulate real-world scenarios where fashion images might be obstructed or incomplete. The diversity in occlusion types allows us to explore the impact of different occlusion characteristics on FIR model performance.

Our analysis goes beyond evaluating model performance and delves into the underlying causes of distribution shift. We explore how different types and sizes of occlusions influence the model's ability to retrieve relevant fashion images. The performance experiences a significant decline as the occlusion ratio increases, with the lowest performance and most distribution shift observed particularly in cases of center occlusion. With the model fine-tuned on the occluded dataset, we observed a more closely aligned distribution and successfully maintained performance to a considerable degree.

2 Related Works

2.1 Unimodal Image Retrieval

Unimodal image retrieval refers to the task of finding images that are relevant to a query image from a large database. The common approach involves two steps – first, using an image encoder to obtain embeddings of query and database images, and second, utilizing a ranking model to rank all database embeddings relative to the query embedding. For the task of finding images from a database containing an object present in the query image, several approaches have been explored. The same object may appear differently in different real-world images, and the embeddings of all the will lie on some manifold in the feature space.

To extract embeddings, CNN-based [15,23,40], or Transformer-based [10,29,36] backbones are often used. In the ranking stage, an appropriate distance measure is used to rank images. The choice of distance metrics depends on the nature of the embedding space and manifold, such as cosine, Euclidean [4,26], and hyperbolic distance measures [11]. Approaches such as query expansion [3,6,12,13,34] and spatial verification [1,32] have been explored to improve the overall performance of image retrieval models. Image retrieval models are typically either trained in a supervised metric learning setting [2,8,14,33,35,38,39,47], or unsupervised metric learning setting [2,16–18,20,46].

Specifically in the domain of fashion image search and retrieval (FIR), deep learning-based approaches have been previously explored [27,30,45,48]. The DeepFashion [24] dataset provides a collection of datasets that are suitable for various real-world scenarios. A subset of DeepFashion, called the InShop Clothes Retrieval dataset, is a smaller dataset useful for FIR research.

2.2 Image Retrieval with Occlusion

Occluded image retrieval refers to the specific case of image retrieval where the object to be searched for in the query image has been occluded by another object, and only part of the object can be seen. Previous occluded image retrieval focused on face image dataset as it can be used for the face identification of criminal suspects from the facial images of CCTV cameras or mobile devices.

Park et al. [31] presented a partially occluded facial image retrieval method based on a similarity measurement for forensic applications. It suggested the novel occluded image retrieval algorithm measuring the similarity based on Scale Invariant Feature Transform (SIFT) matching between normal gallery images and occluded probe images.

Tu et al. [43] proposed two variational autoencoder(VAE)-based networks for predicting a set of unoccluded images that are well matched to the input occluded face, namely, the geometric prediction model and face recovery model.

Mask-FPAN [19] proposed a de-occlusion module that learns to parse occluded faces in a semi-supervised manner regarding face landmark localization, face occlusion estimations and detect head poses.

In the case of fashion image search, in real-world settings it is common for fashion items to be occluded by other objects in view. To the best of our knowledge, there does not exist a dataset of occluded fashion images to test various models' robustness to occluded fashion images in the space of search and retrieval.

3 Occluded Fashion Image Dataset

Our main focus is on the real-world occlusions found in clothing websites, where the query images are provided by the users. User-uploaded images are likely to contain objects or regions that are partially occluded by other objects in their foreground. To incorporate such occlusions into our experimental setup, we construct datasets and evaluate how retrieval models perform with those query images.

3.1 Problem Definition

To evaluate the robustness of an image retrieval model, we utilize a query set $D_{q_{occ}}$ composed of occluded fashion images and a gallery set D_g composed of non-occluded images. Occluded images for $D_{q_{occ}}$ are generated with the occlusion generation process specified in 3.2. The retrieval process can be formulated as follows:

$$\begin{gathered} \mathrm{R}(x_q) = \min_{x_g^i \in D_g} \mathrm{K} \;\; d(x_q, x_g^i), \\ \text{where } x_q \in D_{q_{occ}}. \end{gathered} \tag{1}$$

Given a query image x_q, the retrieval system computes distance d between x_q and all images $x_g^i (1 \leq i \leq |D_g|)$ from the gallery set, and retrieves K images with minimum distance. The retrieved images are expected to have the same items as the query image.

3.2 Occlusion Generation Process

We applied different strategies to generate occluded images. There are 24 different types of occlusions, varying in the colors, sizes, and positions of the occlusion mask.

For ***colors***, the mask is either black or real-colored objects. This is to investigate both cases of occlusion, where it is naively masked (black) or masked in a way it is more likely to be in the real world. We set the two scenarios which occlusion object has its own pattern and color, and the other that doesn't have pattern and color and has the same color with the background. The black occlusion object indicates the situation that doesn't have any pattern and color and has the same color as background, and the color occlusion object indicates that the occlusion object has its own pattern and color that distinguishes it from the real clothes.

For ***sizes***, we pre-define the ratio (e.g., 5%, 10%, 20%) of how many pixels the mask will occupy out of the entire image, and apply an algorithm shown in Algorithm 1 to adjust the size of the mask.

For ***positions***, we give four options; top, bottom, center, and random. That is, we locate the mask either at the top, bottom, center or a random position of the original image. The algorithm for this process is shown in Algorithm 2.

3.3 IS-Occ Dataset

With the generation process specified in 3.2, we construct a new occluded dataset that is built upon InShop Clothes Retrieval dataset [24]. On top of the original dataset, we build 24 new sets of datasets with different types of occlusions and name it **IS-Occ Dataset** (**I**n**S**hop-**Occ**luded **Dataset**). Each subset contains images occluded with a mask with specific color type, size, and position. We only employ these occluded images as the query images, while gallery images remain as original images.

InShop Clothes Retrieval dataset InShop Clothes Retrieval dataset is a subset of DeepFashion database [24]. This subset encompasses considerable pose and scale variations, along with diverse clothing items and comprehensive annotations. The dataset comprises 7,982 clothing items, 52,712 in-shop clothes images, and approximately 200,000 cross-pose/scale pairs.

Microsoft Common Objects in Context (MSCOCO) Image occlusion is carried out by applying additional object masks from the MSCOCO object detection 2017 dataset [21]. A curated ensemble of 36 distinct objects was meticulously chosen to compose a mask set inclusive of entities such as bag, umbrella and person. In addition to black object occlusion, we extracted 128 common objects across 20 categories (e.g., backpack, tie, and cell phone) from the colored common MSCOCO dataset, following [44].

Example of Occlusion Objects We selected two kinds of occlusion objects; *black* occlusion objects and *color* occlusion objects. As shown in Fig.2, we chose common objects in the real world that are often found with clothes such as bag and tie from MSCOCO [21] dataset.

Applying Occlusions in Different Manner Given masks are applied in several different manners as mentioned in 3.2, and we use abbreviations to refer to the specific occlusion type throughout the paper. The occluded object color types include black and real-colored, and for the abbreviation, we use 'Black' and 'Color', respectively. The occluded object size varies from 5%, 10%, and

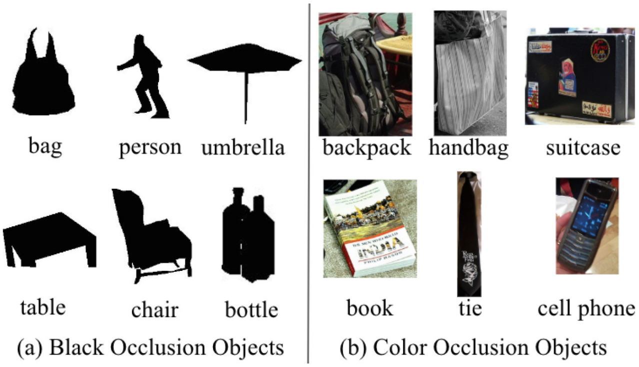


Fig. 2. Example of occlusion objects We chose common objects in the real world that are often found with clothes from MSCOCO dataset.

20% of the total image pixels, and the location varies from top, center, bottom and random location of the image. We put together the first letter of the location and the percentage of the mask size together, e.g., ‘T10’, for the abbreviation. The type of occlusion object is randomly determined for all images. The sample images from the dataset are shown in Fig.3.

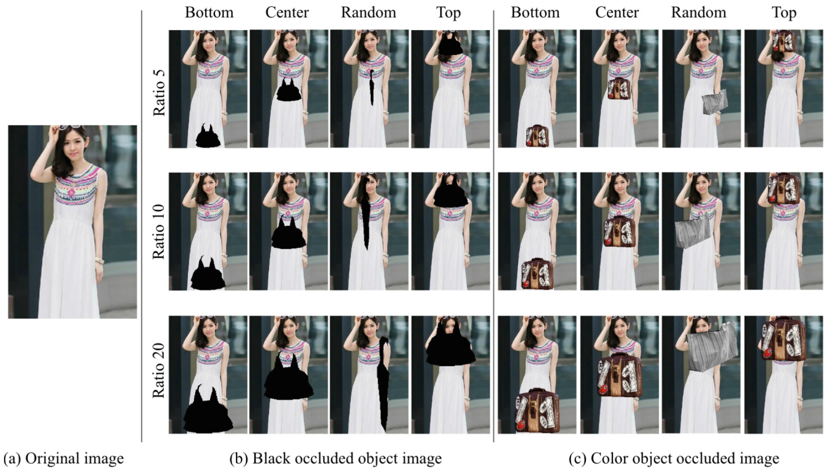


Fig. 3. Example of image with occluded objects of different ratios and locations We experimented with 3 different ratios (5%, 10%, 20%) and 4 different locations (Top, Center, Bottom, Random) and 2 different colors (Black, Color) of the random occluded objects on the Inshop dataset. In this example, for the black setting, bag and umbrella are used as occlusion objects and for the color setting, suitcase and handbag are used.

Algorithm 1: Adjust the ***size*** of the mask

```
Function AdjustMaskSize($I$, $M$, $\rho$):
    input  : original image $I$, mask $M$, masking ratio $\rho \in \{5, 10, 20\}$
    output: resized mask $M$
    $h, w \leftarrow \texttt{size}(I)$
    $M \leftarrow \texttt{resize}(M, (h, w))$
    $\rho_{curr} \leftarrow \frac{\texttt{num\_mask\_pixels}(M)}{h * w}$
    $s \leftarrow \sqrt{\frac{\rho}{\rho_{curr}}}$
    $h_{dest}, w_{dest} \leftarrow \lfloor h * s \rfloor, \lfloor w * s \rfloor$
    $M \leftarrow \texttt{resize}(M, (h_{dest}, w_{dest}))$
    return $M$
```

Algorithm 2: Apply the mask to a certain ***position***

```
Function LocateMask($I$, $M$, $pos$):
    input  : original image $I$, (resized) mask $M$, mask position pos $\in$ {top, bottom,
             center, random}
    output: masked image $I_M$
    $I_M \leftarrow I$
    $h, w \leftarrow \texttt{size}(I)$
    $h_M, w_M \leftarrow \texttt{size}(M)$
    $x_c, y_c \leftarrow \lfloor \frac{w}{2} \rfloor, \lfloor \frac{h}{2} \rfloor$
    $x_{mc}, y_{mc} \leftarrow \lfloor \frac{w_M}{2} \rfloor, \lfloor \frac{h_M}{2} \rfloor$
    if $pos$ = $center$ then
        $x_{start} \leftarrow x_c - x_{mc}$
        $y_{start} \leftarrow y_c - y_{mc}$
    else if $pos$ = $top$ then
        $x_{start} \leftarrow \max(0, x_c - x_{mc})$
        $y_{start} \leftarrow 0$
    else if $pos$ = $bottom$ then
        $x_{start} \leftarrow \max(0, x_c - x_{mc})$
        $y_{start} \leftarrow \max(0, h - h_M)$
    else if $pos$ = $random$ then
        $x_{start}, y_{start} \leftarrow \texttt{random}(2)$
        $x_{start} \leftarrow \min(x_{start}, \max(0, w))$
        $y_{start} \leftarrow \min(y_{start}, \max(0, h))$
    if $x_{start} + w_M \leq w$ then
        $x_{end} \leftarrow x_{start} + w_M$
    else
        $x_{end} \leftarrow w$
    if $y_{start} + h_M \leq h$ then
        $y_{end} \leftarrow y_{start} + h_M$
    else
        $y_{end} \leftarrow h$
    for $i \leftarrow 0$ to $w_M - 1$ do
        for $j \leftarrow 0$ to $h_M - 1$ do
            $I_M[x_{start} + i, y_{start} + j] = M[i, j]$
        end
    end
    return $I_M$
```

Table 1. Results of Black and Color Occluded Object Image Retrieval. We experimented four models with 3 different ratios(5%, 10%, 20%) and 4 different locations (Top, Center, Bottom, Random) and 2 different colors (Black, Real-colored) of the occluded image on the Inshop dataset. **ZS, FT, FT*** indicate zero-shot, fine-tuned on original images, and fine-tuned on 5% occluded images, respectively. Numbers in **bold** indicate the largest performance drop under each experimental setting.

			Black											
		Original	T5	T10	T20	C5	C10	C20	B5	B10	B20	R5	R10	R20
Unicom	ZS	78.86	60.87	50.40	43.56	55.20	42.19	**26.13**	69.33	62.81	54.54	66.32	61.73	55.01
	FT	95.56	92.29	86.66	76.42	91.56	85.31	**71.70**	94.34	92.64	89.04	93.96	92.29	87.27
	FT*	95.68	92.83	87.34	77.75	91.48	85.56	**72.50**	94.44	93.06	89.58	94.27	92.63	87.54
Hyp_ViT	ZS	43.19	26.59	16.68	8.70	23.94	13.04	**4.54**	29.66	18.38	9.88	31.21	21.49	12.84
	FT	92.40	87.02	78.91	64.57	83.42	72.40	**53.43**	89.47	85.95	78.00	88.73	84.93	75.47
	FT*	92.41	90.02	86.70	79.88	89.98	86.81	**78.13**	91.90	90.74	87.95	91.45	90.29	86.64
Hyp_Dino	ZS	46.09	30.23	19.49	11.13	19.96	10.96	**5.14**	26.09	13.86	8.26	27.12	16.73	10.50
	FT	91.19	85.14	77.09	64.43	80.12	69.24	**54.13**	88.79	85.36	76.92	85.91	80.69	69.05
	FT*	89.25	87.79	83.44	76.04	87.99	82.92	**72.24**	90.29	88.67	85.18	89.63	87.90	81.45
Hyp_DeiT	ZS	37.95	16.07	8.05	3.89	12.89	6.57	**2.89**	14.31	7.31	3.97	11.87	6.10	3.55
	FT	91.12	83.85	76.23	63.90	80.17	70.05	**54.69**	88.16	84.89	78.08	86.40	81.37	70.80
	FT*	88.39	87.00	82.66	74.19	87.14	82.40	**71.24**	88.98	87.38	83.29	88.74	87.14	81.17
			Color											
		Original	T5	T10	T20	C5	C10	C20	B5	B10	B20	R5	R10	R20
Unicom	ZS	78.86	55.65	28.87	8.51	44.66	21.65	**6.59**	61.42	37.22	12.06	58.40	44.26	16.20
	FT	95.56	92.43	83.55	64.02	91.43	83.58	**59.11**	94.20	91.74	80.63	93.58	91.45	79.93
	FT*	95.68	93.01	84.54	65.17	91.60	84.05	**59.46**	94.52	92.12	81.59	93.71	91.68	80.34
Hyp_ViT	ZS	43.19	27.11	16.03	5.35	22.82	11.52	**3.71**	28.86	17.34	6.98	28.79	21.77	9.06
	FT	92.40	87.74	78.16	57.27	84.46	72.82	**46.51**	90.71	86.02	69.94	88.87	85.42	67.96
	FT*	92.41	90.82	87.82	**78.82**	91.15	88.93	80.48	91.91	91.23	87.15	91.73	90.98	86.81
Hyp_Dino	ZS	46.09	15.00	3.35	1.51	3.91	1.53	1.29	12.86	3.47	1.41	9.95	3.64	**1.14**
	FT	91.19	77.49	50.06	14.56	58.13	26.60	**7.08**	83.30	61.90	18.33	76.29	59.09	18.45
	FT*	89.25	88.90	83.90	71.70	88.85	85.17	**70.87**	90.42	88.48	80.71	90.34	88.99	79.91
Hyp_DeiT	ZS	37.95	5.93	2.06	1.29	4.47	1.66	1.16	5.95	2.29	1.24	5.38	2.43	**0.88**
	FT	91.12	79.50	56.54	21.01	69.54	41.05	**13.25**	85.26	73.54	33.57	81.74	71.33	33.00
	FT*	88.39	87.62	82.94	71.51	88.00	84.30	**70.59**	89.15	87.47	79.58	89.24	88.19	79.03

4 Experiments

We run experiments to investigate how different types of occlusions affect the retrieval performance. We evaluate the models of three settings; zero-shot, fine-tuned on original images, and fine-tuned on Color-C5 occluded images. For zero-shot experiments, we directly use the pre-trained models. For evaluation with fine-tuned models, the pre-trained models are fine-tuned either on the original In-Shop dataset or the corresponding 5% occluded dataset during query phase. The overall results are shown in Table 1.

4.1 Implementation Details

We used pre-trained visual backbone encoders from Unicom [2], Hyperbolic ViT [11], Hyperbolic Dino [5], and Hyperbolic DeiT [42] as our backbone model. The

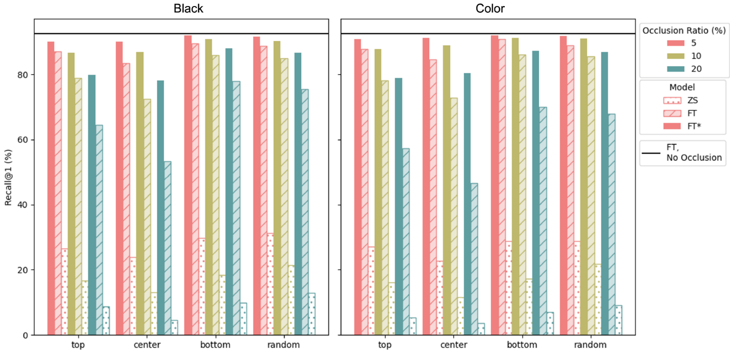


Fig. 4. Retrieval performance of Hyp_ViT in various setups. Different bar colors represent different occlusion sizes, and filled patterns denote the model used. The horizontal axis shows occlusion positions. The left plot shows results with black mask occlusions, and the right plot shows results with colored object occlusions.

training is conducted on 1 NVIDIA T4 Tensor Core GPU. We use AdamW [25] as the optimizer with an initial learning rate of 0.001, and a weight decay of 0.05. We used margin-based softmax loss, ArcFace [9], for both pre-training and image retrieval tasks.

4.2 Ablations on Occlusion Strategies

We carry out the ablation studies on two occlusion types (black and real-object) with four different locations (top, center, bottom, random) and three different occlusion ratios (5%, 10%, 20%). To see which occlusion affects the retrieval performance the most, we conduct image retrieval with entire sets of IS-Occ datasets.

As shown in Table 1 and Fig. 4, different types and sizes of occlusions bring different extents of performance drop. Among 24 occluded datasets we created, Black-B5 had the least impact, especially when it was tested on fine-tuned models. On the other hand, Color-C20 had the most significant impact on the performance.

We found out that ***real-colored objects*** rather than black masks more confuse the model and lead to a drastic performance drop. One possible reason might be the colors and patterns that the real-colored objects contain. Since real-colored objects often contain irregular patterns unlike mono-colored masks, it is more likely to confuse them with the clothes that the model has to concentrate on.

In terms of ***sizes***, we found that the model gives poor performance with the bigger masks. This could be seen obvious, since the bigger mask has more probability of covering the entire clothes or a large portion of them.

Lastly, among different ***positions*** of occlusions, we find out that masking the center part of the image affects the model to degrade in performance the most. Since the In-shop dataset originally contains clean images with the clothes mostly in the center, this could be thought as a consequence of masking the core part of the clothes. Among the top and bottom, masking the top part of the image usually showed a larger performance drop even with 5% of occlusion. By this, we carefully assume that masking the facial part of the image affects the performance of fashion image retrieval. We leave further analysis on this observation as future work.

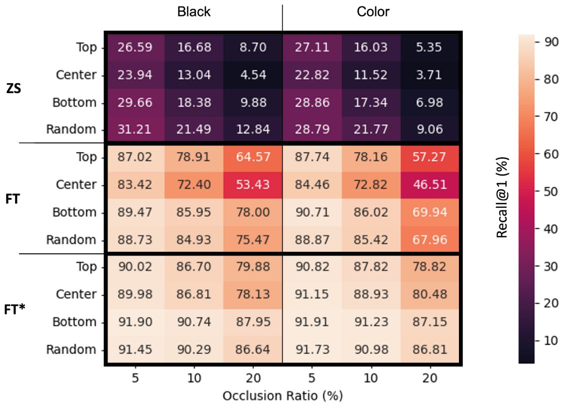


Fig. 5. Retrieval performances in R@1 (%), for each occlusion scenario.

Fig. 5 clearly summarizes the result of our ablation with Hyp_ViT. Zero-shot inference results (**ZS**) are represented in dark colors, indicating their poor performance, particularly with big occlusions. While fine-tuned models—**FT** and **FT***—exhibit much better performance on occluded images compared to **ZS**, it is especially prone to occlusions that are larger in size and placed in the top or center of the image. Robustness against occlusion sizes and positions is largely improved in **FT***, the model trained on occluded images.

4.3 Qualitative Results

The examples of retrieved images with Hyp_ViT is shown in Fig 6. As shown in the figure, zero-shot model suffers from retrieving correct gallery images in both original and occluded dataset. In the case of model fine-tuned on the original dataset, it performs well when the original query image is given, but it suffers from retrieving correct gallery images when occluded query image is given. The results show that the model fine-tuned on IS-Occ dataset is more robust to the occlusion. Qualitative results for Hyp_Dino and Hyp_DeiT are provided in the supplementary material.

5 Analysis

5.1 Distribution Shift of Occluded Datasets

In this section, we analyze the results in terms of the distribution shift caused by occlusion. Since the zero-shot model was only exposed to complete images during the training phase, it would not have had the chance to learn how to generalize to occluded images. This distribution shift between train and test data is known to degrade the test performance.

To investigate how our occluded data has shifted in distribution, we use PCA(Principal Component Analaysis) to visualize the feature space. We visual-

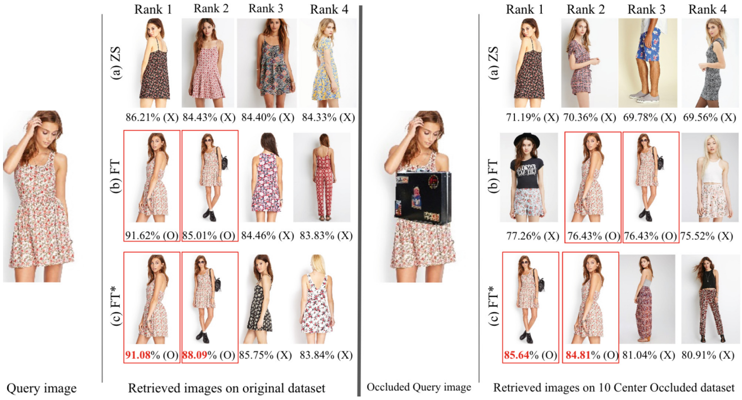

Rank 1
Rank 2
Rank 3
Rank 4
(a) ZS
86.21% (X) 84.43% (X) 84.40% (X) 84.33% (X)
(b) FT
91.62% (O) 85.01% (O) 84.46% (X) 83.83% (X)
(c) FT*
91.08% (O) 88.09% (O) 85.75% (X) 83.84% (X)
Query image
Retrieved images on original dataset
Rank 1
Rank 2
Rank 3
Rank 4
(a) ZS
71.19% (X) 70.36% (X) 69.78% (X) 69.56% (X)
(b) FT
77.26% (X) 76.43% (O) 76.43% (O) 75.52% (X)
(c) FT*
85.64% (O) 84.81% (O) 81.04% (X) 80.91% (X)
Occluded Query image
Retrieved images on 10 Center Occluded dataset

Fig. 6. Retrieved images with non-occluded (left) and occluded (right) query image. Each row indicates the top 4 ranked retrieved images and their confidence scores for each model—(a) **ZS**(zero-shot), (b) **FT**(fine-tuned on non-occluded images), and (c) **FT***(fine-tuned on occluded images). Correctly retrieved images are represented with red edges.

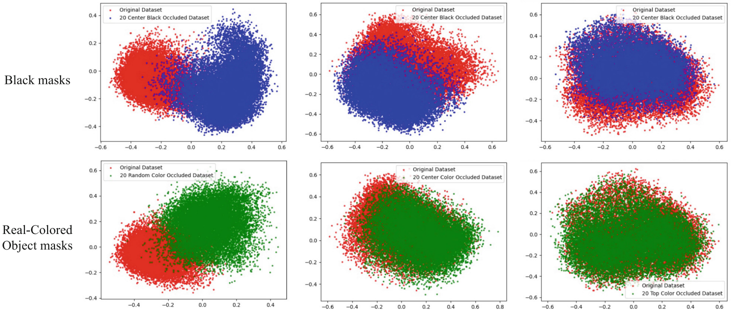
Zero-Shot
Fine-tuned on Original Images
Fine-tuned on 5% Occluded Images


Fig. 7. PCA visualization of Hyp_ViT feature representation on original images and occluded dataset The distribution of features from zero-shot, original dataset finetuned, Black-20 and Color-20 occluded dataset finetuned Hyp_ViT model. Extracted features of the original dataset are in red, those of the black-occluded dataset are in blue, and those with real-colored-occlusions are in green.

Table 2. CLIP score of IS-Occ dataset CLIP score between the representations of DINOv2 of original image and occluded image. *Mean* and *Std.* refers to the average and standard deviation of the clip scores of all the images, respectively. Numbers in **bold** indicate the largest performance drop under each occlusion setting.

	T5	T10	T20	C5	C10	C20
Mean	0.858	0.840	0.804	0.860	0.845	0.809
Std.	0.050	0.050	0.048	0.051	0.051	0.049
	B5	B10	B20	R5	R10	R20
Mean	0.859	0.844	0.808	0.857	0.836	**0.803**
Std.	0.052	0.051	0.046	0.051	0.051	0.047

ize output features of original images and those of occluded images from three evaluation settings, to see if the fine-tuning process affects the distribution gap between two different sets of images.

Fig.7 shows the result of dimensionality reduction of the feature space with query instances. Different types of occlusions yield different extents of shifts. The result shows the existence of a distribution gap between the features of original images and occluded images. Among the three models, the model without fine-tuning on Inshop dataset yields the biggest shift, and also the poorest performance. The distributions of original and occluded images become closer along with better performance, as they are fine-tuned on Inshop dataset. Finally, fine-tuning the model on the occluded dataset with Color-20 gives distributions that almost overlap, along with its robust performance observed in Table 1.

To this end, we assume the main reason the model finds it hard to retrieve images with occluded query images, is due to such distribution shift. Closing the gap between representations of non-occluded and occluded data helps the model to generalize with other types of occlusions as well, eventually achieving robustness against occlusions.

5.2 CLIP score of Image Embeddings

We also compared the CLIP score of the image representation to explore the similarity of images with different occlusion ratios and positions. CLIP score is often used to calculate the similarity between two representation. We used DINOv2.CLIP [28] model to first extract the feature representation of the original image and occluded image and then calculate cosine similarity between two representations.

As shown in Table 2, as occlusion ratio gets bigger, DINOv2's clip score drops. It shows that the model suffers from obtaining proper representation when the image is occluded. Referring to Fig.7, the model fine-tuned on our IS-Occ dataset makes the model robust to occlusion when the model without fine-tuning suffers from perceiving the original image and occluded image are the same.

6 Conclusion

This project tackles the challenge of fashion image retrieval under an occlusion setting of the In-Shop Fashion Image Retrieval dataset. We introduced occlusions from the MSCOCO dataset to simulate realistic scenarios of partially obscured clothing items. Results show that state-of-the-art image retrieval models experience performance drop with occluded images. Fine-tuning on our IS-Occ dataset with 5% occlusion significantly enhances model robustness, reducing the performance drop. Principal Component Analysis of feature space distribution reveals a substantial gap in models not fine-tuned on In-Shop, leading to poor occlusion performance. Conversely, fine-tuned models on a 5% occluded dataset exhibit nearly overlapping distributions and robust performance, indicating targeted fine-tuning minimizes distribution shifts.

Despite our findings from exploring different types of occlusions, there still remains diverse realistic scenarios that we could not fully cover. This paper is vital in raising the problem caused by occlusions in fashion image retrieval, by employing a basic set of synthetic occlusions. However, more realistic scenarios, such as an obstructed view of the target item due to other people or a moving vehicle, should be considered for future work. Furthermore, we would like to explore the integration of a reconstruction approach, such as in-painting models and adding adapters in fine-tuning.

Acknowledgement. This work was supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government(MSIT) (RS-2022-00143911, AI Excellence Global Innovative Leader Education Program)

References

1. An, G., Huo, Y., Yoon, S.E.: Hypergraph propagation and community selection for objects retrieval. Adv. Neural. Inf. Process. Syst. **34**, 3596–3608 (2021)
2. An, X., Deng, J., Yang, K., Li, J., Feng, Z., Guo, J., Yang, J., Liu, T.: Unicom: Universal and compact representation learning for image retrieval. In: The Eleventh International Conference on Learning Representations (2022)
3. Arandjelović, R., Zisserman, A.: Three things everyone should know to improve object retrieval. In: 2012 IEEE conference on computer vision and pattern recognition. pp. 2911–2918. IEEE (2012)
4. Babenko, A., Lempitsky, V.: Efficient indexing of billion-scale datasets of deep descriptors. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 2055–2063 (2016)
5. Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., Joulin, A.: Emerging properties in self-supervised vision transformers (2021), https://arxiv.org/abs/2104.14294
6. Chum, O., Philbin, J., Sivic, J., Isard, M., Zisserman, A.: Total recall: Automatic query expansion with a generative feature model for object retrieval. In: 2007 IEEE 11th International Conference on Computer Vision. pp. 1–8. IEEE (2007)

7. Corbiere, C., Ben-Younes, H., Ramé, A., Ollion, C.: Leveraging weakly annotated data for fashion image retrieval and label prediction. In: Proceedings of the IEEE international conference on computer vision workshops. pp. 2268–2274 (2017)
8. Deng, J., Guo, J., Zafeiriou, S.: Arcface: Additive angular margin loss for deep face recognition. 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) pp. 4685–4694 (2018), https://api.semanticscholar.org/CorpusID:8923541
9. Deng, J., Guo, J., Xue, N., Zafeiriou, S.: Arcface: Additive angular margin loss for deep face recognition. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 4690–4699 (2019)
10. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. In: International Conference on Learning Representations (2020)
11. Ermolov, A., Mirvakhabova, L., Khrulkov, V., Sebe, N., Oseledets, I.: Hyperbolic vision transformers: Combining improvements in metric learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7409–7419 (2022)
12. Gordo, A., Almazan, J., Revaud, J., Larlus, D.: End-to-end learning of deep visual representations for image retrieval. Int. J. Comput. Vision **124**(2), 237–254 (2017)
13. Gordo, A., Radenovic, F., Berg, T.: Attention-based query expansion learning. In: European Conference on Computer Vision. pp. 172–188. Springer (2020)
14. Hadsell, R., Chopra, S., LeCun, Y.: Dimensionality reduction by learning an invariant mapping. In: 2006 IEEE computer society conference on computer vision and pattern recognition (CVPR'06). vol. 2, pp. 1735–1742. IEEE (2006)
15. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
16. Iscen, A., Tolias, G., Avrithis, Y., Chum, O.: Mining on manifolds: Metric learning without labels. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition pp. 7642–7651 (2018), https://api.semanticscholar.org/CorpusID:4466042
17. Kan, S., Cen, Y., Li, Y., Mladenovic, V., He, Z.: Relative order analysis and optimization for unsupervised deep metric learning. 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) pp. 13994–14003 (2021), https://api.semanticscholar.org/CorpusID:235691639
18. Kim, S., Kim, D., Cho, M., Kwak, S.: Self-taught metric learning without labels. 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) pp. 7421–7431 (2022), https://api.semanticscholar.org/CorpusID:248512812
19. Li, L., Zhang, T., Kang, Z., Jiang, X.: Mask-fpan: Semi-supervised face parsing in the wild with de-occlusion and uv gan. Computers & Graphics **116**, 185–193 (2023)
20. Li, Y., Kan, S., He, Z.: Unsupervised deep metric learning with transformed attention consistency and contrastive clustering loss. ArXiv **abs/2008.04378** (2020), https://api.semanticscholar.org/CorpusID:221095511
21. Lin, T.Y., Maire, M., Belongie, S., Bourdev, L., Girshick, R., Hays, J., Perona, P., Ramanan, D., Zitnick, C.L., Dollár, P.: Microsoft coco: Common objects in context (2015)

22. Lin, Y.L., Tran, S., Davis, L.S.: Fashion outfit complementary item retrieval. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 3311–3319 (2020)
23. Liu, Z., Mao, H., Wu, C.Y., Feichtenhofer, C., Darrell, T., Xie, S.: A convnet for the 2020s. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 11976–11986 (2022)
24. Liu, Z., Luo, P., Qiu, S., Wang, X., Tang, X.: Deepfashion: Powering robust clothes recognition and retrieval with rich annotations. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2016)
25. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101 (2017)
26. Muja, M., Lowe, D.G.: Scalable nearest neighbor algorithms for high dimensional data. IEEE Trans. Pattern Anal. Mach. Intell. **36**(11), 2227–2240 (2014)
27. Naka, R., Katsurai, M., Yanagi, K., Goto, R.: Fashion style-aware embeddings for clothing image retrieval. In: Proceedings of the 2022 International Conference on Multimedia Retrieval. pp. 49–53 (2022)
28. Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., et al.: Dinov2: Learning robust visual features without supervision. arXiv preprint arXiv:2304.07193 (2023)
29. Oquab, M., Darcet, T., Moutakanni, T., Vo, H.V., Szafraniec, M., Khalidov, V., Fernandez, P., HAZIZA, D., Massa, F., El-Nouby, A., Assran, M., Ballas, N., Galuba, W., Howes, R., Huang, P.Y., Li, S.W., Misra, I., Rabbat, M., Sharma, V., Synnaeve, G., Xu, H., Jegou, H., Mairal, J., Labatut, P., Joulin, A., Bojanowski, P.: DINOv2: Learning robust visual features without supervision. Transactions on Machine Learning Research (2024), https://openreview.net/forum?id=a68SUt6zFt
30. Park, S., Shin, M., Ham, S., Choe, S., Kang, Y.: Study on fashion image retrieval methods for efficient fashion visual search. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. pp. 0–0 (2019)
31. Park, S., Lee, H., Yoo, J.H., Kim, G., Kim, S., et al.: Partially occluded facial image retrieval based on a similarity measurement. Mathematical Problems in Engineering **2015** (2015)
32. Philbin, J., Zisserman, A.: Object mining using a matching graph on very large image collections. In: 2008 Sixth Indian Conference on Computer Vision, Graphics & Image Processing. pp. 738–745. IEEE (2008)
33. Qian, Q., Shang, L., Sun, B., Hu, J., Li, H., Jin, R.: Softtriple loss: Deep metric learning without triplet sampling. 2019 IEEE/CVF International Conference on Computer Vision (ICCV) pp. 6449–6457 (2019), https://api.semanticscholar.org/CorpusID:202558557
34. Radenović, F., Tolias, G., Chum, O.: Fine-tuning cnn image retrieval with no human annotation. IEEE Trans. Pattern Anal. Mach. Intell. **41**(7), 1655–1668 (2018)
35. Schroff, F., Kalenichenko, D., Philbin, J.: Facenet: A unified embedding for face recognition and clustering. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 815–823 (2015)
36. Shaker, A.M., Maaz, M., Rasheed, H.A., Khan, S., Yang, M., Khan, F.S.: Swiftformer: Efficient additive attention for transformer-based real-time mobile vision applications. 2023 IEEE/CVF International Conference on Computer Vision (ICCV) pp. 17379–17390 (2023), https://api.semanticscholar.org/CorpusID:257766532

37. Shiau, R., Wu, H.Y., Kim, E., Du, Y.L., Guo, A., Zhang, Z., Li, E., Gu, K., Rosenberg, C., Zhai, A.: Shop the look: Building a large scale visual shopping system at pinterest. In: Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. pp. 3203–3212 (2020)
38. Sohn, K.: Improved deep metric learning with multi-class n-pair loss objective. Advances in neural information processing systems **29** (2016)
39. Sun, Y., Cheng, C., Zhang, Y., Zhang, C., Zheng, L., Wang, Z., Wei, Y.: Circle loss: A unified perspective of pair similarity optimization. 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) pp. 6397–6406 (2020), https://api.semanticscholar.org/CorpusID:211296865
40. Tan, M., Le, Q.: EfficientNet: Rethinking model scaling for convolutional neural networks. In: Chaudhuri, K., Salakhutdinov, R. (eds.) Proceedings of the 36th International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 97, pp. 6105–6114. PMLR (09–15 Jun 2019), https://proceedings.mlr.press/v97/tan19a.html
41. Tian, Y., Newsam, S., Boakye, K.: Fashion image retrieval with text feedback by additive attention compositional learning. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 1011–1021 (2023)
42. Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A., Jégou, H.: Training data-efficient image transformers & distillation through attention (2021), https://arxiv.org/abs/2012.12877
43. Tu, C.T., Lee, K.H.: Occluded face recovery by image retrieval. In: 2021 International Symposium on Intelligent Signal Processing and Communication Systems (ISPACS). pp. 1–2. IEEE (2021)
44. Voo, K.T., Jiang, L., Loy, C.C.: Delving into high-quality synthetic face occlusion segmentation datasets. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4711–4720 (2022)
45. Yan, C., Yan, K., Zhang, Y., Wan, Y., Zhu, D.: Attribute-guided fashion image retrieval by iterative similarity learning. In: 2022 IEEE International Conference on Multimedia and Expo (ICME). pp. 1–6. IEEE (2022)
46. Yan, J., Luo, L., Deng, C., Huang, H.: Unsupervised hyperbolic metric learning. 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) pp. 12460–12469 (2021), https://api.semanticscholar.org/CorpusID:235693274
47. Zhai, A., Wu, H.Y.: Classification is a strong baseline for deep metric learning. In: British Machine Vision Conference (2018), https://api.semanticscholar.org/CorpusID:199442350
48. Zhu, J., Huang, H., Deng, Q.: Fashion image retrieval with multi-granular alignment. arXiv preprint arXiv:2302.08902 (2023)

Oracle Bone Inscription Image Retrieval Based on Improved ResNet Network

Jun Ding[1], Jiaoyan Wang[1], Alimjan Aysa[1,2], Xuebin Xu[1,2], and Kurban Ubul[1,2,3](✉)

[1] School of Computer Science and Technology, Xinjiang University, Ürümqi 830046, China
kurbanu@xju.edu.cn

[2] Xinjiang Multilingual Information Technology Key Laboratory, Xinjiang University, Ürümqi 830046, China

[3] Joint International Research Laboratory of Silk Road Multilingual Cognitive Computing, Xinjiang University, Urumqi 830046, China

Abstract. Research on oracle bone inscription image retrieval is important for applications in academic and cultural heritage areas. The current oracle bone dataset faces problems such as the low similarity between the same category, the high similarity between the different categories, and imbalanced sample distribution. In addition, due to the complex background of oracle bone images, existing network models have certain limitations in extracting image features. To address these challenges, this study first adopts a Siamese network-based image retrieval method to learn feature representations of similar and dissimilar images. Subsequently, the existing dataset was partitioned, providing a practical and usable retrieval dataset for the oracle bone image retrieval field. Finally, an improved network model based on ResNet is proposed and integrated into the Siamese network framework. The model achieves the highest retrieval MP and MAP values of 83.26% and 90.68%, respectively, which is better than the current research.

Keywords: Feature extraction · Image retrieval · Oracle bone inscription · Siamese network · Sample imbalance

1 Introduction

As the earliest Chinese writing with a mature language system, oracle bone inscriptions are widely regarded as an important source for the study of ancient society, history, and culture. oracle bone inscription image retrieval aims to retrieve images similar to the query image from a large number of databases and sort them according to similarity. In contrast to the retrieval of modern Chinese character images [1], there has been no systematic research on the retrieval of oracle bone inscription images in China. First, there is the problem of low similarity between the same class and high similarity between the different classes,

A. Antonacopoulos et al. (Eds.): ICPR 2024, LNCS 15321, pp. 47–62, 2025.
https://doi.org/10.1007/978-3-031-78305-0_4

which makes oracle bone inscription image retrieval much more difficult. Second, due to historical reasons, most oracle bones have severe damage and a large number of missing image structures, which causes a serious sample imbalance problem. Finally, the background of oracle bone inscription images is complex, and the existing network has insufficient feature extraction capability for oracle bone images.

Oracle bone inscription has a large amount of data but lacks unified standards. To this end, Huang et.al [2] created a large-scale oracle bone image dataset OBC306, providing the first publicly available real image dataset for oracle bone image retrieval research. Oracle bone inscription image retrieval consists of two main stages: feature extraction and similarity calculation [3]. In the feature extraction, the image is converted into feature vectors to represent the image content. In the similarity calculation stage, the algorithm evaluates the similarity between the query image and each image in the database, usually by comparing the feature vectors between images or using specific similarity measurement methods. The application of oracle bone script image retrieval helps experts and scholars quickly find images that are similar to the input image, thereby aiding in the deciphering and interpretation of oracle bone inscription. However, at present, there are relatively few studies on oracle bone image retrieval at home and abroad, and this field still belongs to the cold field. Xiong et al [4] proposed a new solution strategy to improve the completeness and accuracy check of literature search in the field of oracle literature by proposing an ontology-based comprehensive search strategy for oracle literature and establishing an oracle literature ontology and an optimized search platform. Oracle image retrieval was initially focused on copying Oracle bone inscriptions. For example, Lin [5] used crawler software to obtain more than 40,000 images of copy oracle bone inscription and used CNN and VLAD methods to generate the representation features of oracle bone images, and its retrieval accuracy on copy oracle bones reached 84.1%. However, the images of the oracle bones have been altered over time and many have been lost, resulting in a small amount of data for some of the samples, and there are more variant characters in the oracle bones, which further increases the difficulty of retrieving oracle bone inscription images. To solve the problem of oracle bone variant character retrieval, Liu et al [6] proposed an image retrieval method combining deep neural network (DNN) and clustering techniques, which improved the retrieval accuracy while reducing the average query time. However, they found that the inter-class similarity has a significant negative effect on the check-all rate of image retrieval. Xu et al [7] constructed an information system for oracle bone information processing (IsOBS), which provides a search function for character and document databases to help users quickly find characters and documents for further research. Although the above studies can solve the problem of difficult oracle bone image retrieval to some extent, most of them are aimed at copying oracle bone images. Due to the complexity of the background of the real oracle bone inscription images, the presence of severe noise, missing characters, broken characters, and long-tailed distribution problems, the retrieval difficulty is significantly higher than that of the copy oracle bone, the research in this area still needs to be explored in depth. In a recent study on oracle bone inscription image retrieval, Yao [8] provided a classification network-based oracle bone image retrieval scheme, which can achieve the retrieval of oracle bone images to a certain extent. However, the retrieval dataset used is separated from the oracle bone inscription dataset, which cannot solve the problem of retrieving unknown oracle bone images. The method lacks a certain degree of generalizability. The contributions of this paper are:

- A Siamese network-based oracle bone inscription image retrieval method was used to achieve a comparison between query images and the image library to be queried. This method can learn feature representations to distinguish similar and dissimilar images, thus achieving efficient retrieval.
- To solve the problem of imbalanced sample distribution, this article divides the OBC306 dataset, providing a usable retrieval dataset for the field of oracle bone inscription image retrieval.
- An improved network model based on ResNet is proposed: using a 7*7 large convolution replaced by three 3*3 small convolutions; designing a skip convolution cascade module structure to pass and fuse features from the shallow layers of the network to the deeper layers of the network.

2 RELATED WORKS

2.1 Image Retrieval and Siamese Network

Image retrieval is a technique used to retrieve and query similar images from a database. It involves extracting image features and calculating similarity measures. This technology is widely used in digital libraries, Internet image search, video surveillance, medical imaging, and other fields. In these fields, it is necessary to quickly and accurately retrieve images similar to query images from a large amount of image data, in order for users to browse, analyze, and apply them. Image retrieval is essentially an operation that measures the similarity between images [9]. In this study, oracle bone inscription image retrieval is considered a metric learning problem, which maps input data to a metric space where similar objects have smaller distances and dissimilar objects have larger distances for image similarity retrieval.

Researchers have proposed similarity measurement learning algorithms, attempting to optimize distance measurement through training data or auxiliary information to improve the performance of image retrieval. A typical metric learning architecture, such as a Siamese network [10] or a triplet network [11], differs from classification in that it only requires object class labels, while for specific objects, the labels must be for each image pair. The distance measurement effect is achieved by training with input matching or mismatched image pairs. Therefore, this dual-branch or triple-branch network architecture is more suitable for metric learning tasks. Siamese neural networks have been widely studied in many fields of computer science, such as Carlevaris et al [12] using a Siamese neural network to learn the features of trained images. Pan. et al [13] consider the Siamese neural network as an effective metric learning method. In addition, Hadsell et al.[14] utilized a siamese neural network to perform dimensionality reduction operations on features. Filip Radenovi' et al.[15] used a Siamese neural network in their paper, the two branches use the same construction method and share parameters, extract the characteristics of the two pictures to be compared, and then use the loss function to calculate the difference. Siamese Networks is a special type of neural network architecture designed to compare the similarity between input samples. Twin neural networks have extensive applications in fields such as image recognition, image search, and image generation in deep learning. They consist of two interrelated neural networks, which typically have the same structure but may have different initialization weights. The core idea is to evaluate the similarity between samples by feeding two input samples into two neural networks and comparing their embedding representations in a high-dimensional feature space [16].

3 Oracle Bone Inscription Image Retrieval Based on Siamese Networks

The Siamese network framework used in this article is shown in Fig.1. This Siamese network framework includes two parts: feature extraction and similarity measurement. Two oracle bone inscription images are used as inputs in both the training and retrieval stages, and the output of the entire framework is the similarity between the two images.

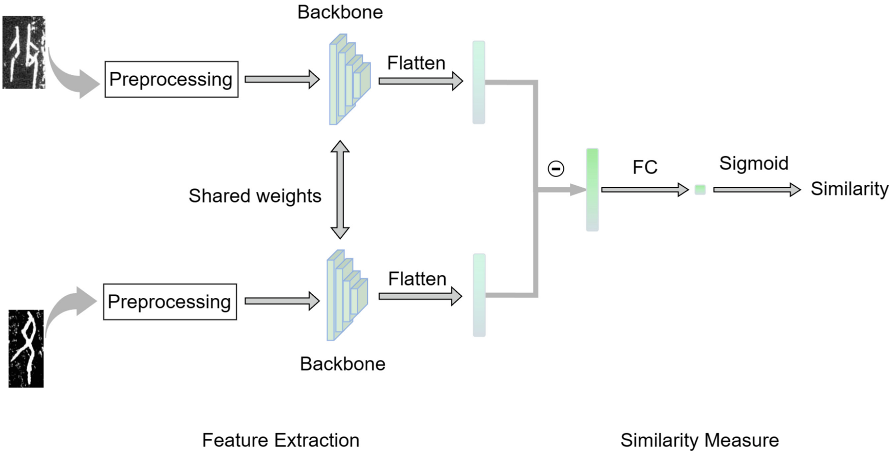


Fig. 1. Image Retrieval of Oracle Bone Inscription Based on Siamese Networks

The goal of Siamese network training is to enable the network to map samples from the same category to nearby areas in high-dimensional space, and to map samples from different categories to distant areas. Fig.1 shows the training process of oracle bone image training, which mainly includes two parts: feature extraction and similarity measurement. In the feature extraction stage: First, input samples from the training stage, including two images from the same or different categories. Secondly, preprocessing operations such as image size adjustment and normalization of pixel values are performed on the input image to ensure consistency Finally, the backbone network is used for feature extraction to obtain feature descriptor vectors for two input images. In the similarity measurement stage: Firstly, the feature descriptor vectors of two images are unfolded, and subtracted, and the absolute value is taken to obtain the feature difference between the two input images. Then, the feature differences are mapped to a hidden vector space through a fully connected layer. Finally, use the activation function sigmoid to convert the hidden vector space into a probability value. This probability value can represent the similarity between two images, with higher values indicating greater similarity between the two input images.

By training the network to minimize the distance between feature vectors of similar samples and maximize the distance between feature vectors of dissimilar samples, the Siamese network can learn feature representations with good discriminative power for similarity. In addition, for the problem of small tail samples, the Siamese networks may provide a solution [17]. Due to the Siamese network requiring a pair of samples for training, it can use categories with only a small number of samples in the dataset for training. For these small sample categories, the sample size is small, training them with negative sample pairs composed of samples from other categories can also teach the characteristics of these categories to a certain extent, which can better utilize small sample information and alleviate the problem of sample imbalance to a certain extent. This makes it a powerful tool for oracle bone inscription image retrieval.

When conducting retrieval, it is necessary to extract features from the query image and the image library to be queried through the same feature extraction network. By sharing network weights, the similarity between the query image vector and each image library vector to be queried is obtained separately. Sort the similarity according to the dictionary and output the similarity from highest to lowest. Finally, based on the search results, a query evaluation is conducted to evaluate the quality of the model.

4 Data Set

This article uses the oracle bone inscription dataset OBC306 constructed by the Key Laboratory of Oracle Bone Information Processing at Anyang Normal University, Ministry of Education. The dataset contains 309551 oracle bone character images, covering 306 types of oracle bone characters. The distribution of category and sample size is shown in Fig.2. From the graph, it can be seen that the OBC306 dataset has a serious long-tail distribution problem, and its severely imbalanced data distribution will cause the model to focus more on categories with larger sample sizes during training while paying insufficient attention to categories with smaller sample sizes. This may make it difficult for the model to correctly learn and recognize categories with smaller sample sizes, and perform better on categories with larger sample sizes.

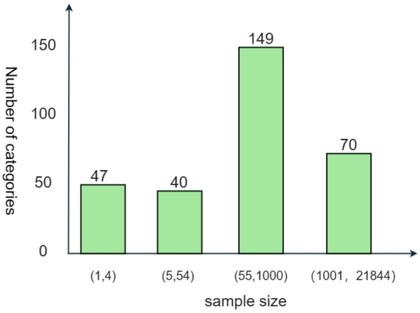


Fig. 2. The category number and sample size distribution of OBC306

In response to the issue of imbalanced sample size distribution in OBC306, many scholars have made different treatments for the tail dataset in OBC306. Mao et al. [18] deleted all classes in the OBC306 dataset that only contained one image, retaining 277 classes, resulting in a total of 309552 images. Zhang et al. [19] selected 241 categories with a large number of samples in the recognition of oracle bone script based on cross-modal deep metric learning, totaling 295466 samples. Each category had a minimum of 16, a maximum of 25898, and an average of 1226 samples. During the research, OBC306 was also divided, and the specific division rules are as follows:

Firstly, from Fig.2, it can be seen that there are 47 categories with less than 5 samples, and these categories contain a total of 74 oracle bone inscription images, with an average of less than two samples per category. Considering that in large-scale datasets such as 300000 levels, these categories with fewer samples may not have a significant impact on overall performance, this study decided to remove these 47 categories from the OBC306 dataset. In the field of oracle bone inscription image retrieval research, both domestically and internationally, there is relatively little research on oracle bone inscription image retrieval, which leads to a lack of suitable datasets for retrieval research. Given the significant head and tail differences in the sample size distribution of the OBC306 dataset, this study decided to further segment the dataset after removing 47 categories. We selected 40 categories at the end, with sample sizes ranging from 5 to 54 images, totaling 1189 images. We named them TOBC40 (Tail-OBC40) and used them as the retrieval dataset. The selection of TOBC40 as the retrieval dataset is mainly based on the following considerations: firstly, it can further alleviate the impact of insufficient tail sample learning; Secondly, provides a fixed retrieval dataset for the field of oracle bone image retrieval; Finally, it can be verified that the proposed method can retrieve unknown oracle bone images. Finally, 219 classes of samples were retained for training, with at least 55 images in each class and a total of 308288 images, named OBC219

5 Backbone

This section will focus on the selection and design of the backbone network in the Siamese network-based image retrieval framework for oracle bone inscription. The effectiveness of the proposed improvements in improving the retrieval performance is demonstrated through extensive experiments and visual analyses.

5.1 Oracle Bone Inscription Network

Residual networks (ResNet) are well known for their unique residual learning architecture, see the study by He [20] for details.ResNet has demonstrated significant performance advantages in image recognition tasks. However, for oracle bone topography images, due to the high similarity between oracle bone characters, some characters with different meanings are almost indistinguishable from each other in terms of contours, and only differ in subtle ways, which leads to a reduction in recognition, thus ResNet's performance on the oracle bone recognition task does not meet the expected results. In this chapter, ResNet18 is used as the backbone network, and the following improvements are made to design a network suitable for Oracle bone inscription image feature extraction: Oracle bone inscription Network (OBINet).OBINet is improved in two aspects on top of ResNet: The first 7*7 convolution of the ResNet network is replaced by three 3*3 convolutions. The skip convolution cascade module structure is designed to transfer and fuse the features from the shallow layers of the network to the deeper layers of the network. The overall structure of the network in this chapter is shown in Fig. 3. In the next 2 subsections, these two improvements will be introduced and the feasibility of the improved network will be experimentally analyzed.

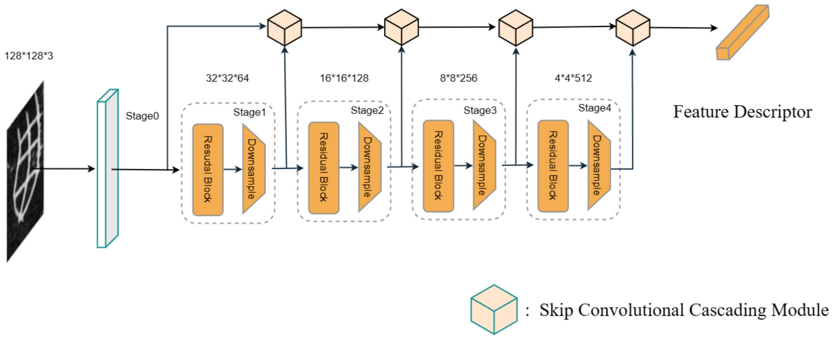


Fig. 3. Overall framework diagram of OBINet

5.2 Design of Replacement Convolution

Oracle bone inscription images are usually created by topography from oracle bones. Most of these image samples have an aspect ratio of about 1:2, with a height of about 120 pixels and a width of about 50 pixels. Therefore, most of the images are tiny images. As a deep learning model, ResNet solves the problem of gradient vanishing in deep network training by constructing residual blocks to achieve a deeper network structure. However, the 7*7 convolutional kernel in a traditional ResNet network may not be the best choice for small-resolution images such as Oracle. Since larger convolutional kernels typically capture a wider range of features, this works well for information-rich images. However, for lower resolution images, it may cause the network to learn too much noise and unnecessary features, which in turn affects the recognition accuracy. Therefore, in this paper, we choose to replace the first convolutional kernel of the stage(0) layer in the ResNet network from 7*7 to 3*3 convolutional kernel and add the ReLU activation function between each 3*3 convolutional layer to increase the nonlinearity. Finally, a batch normalization layer is added between the 3*3 convolutional layers to improve the stability and convergence speed of the model. This modified model is referred to as ResNet-A. This adaptation was originally proposed in Inception-v3 [21] and has been validated in other literature [22]. Such a modification aims to reduce the model complexity while maintaining a certain feature extraction capability, focusing more on extracting key features in lower-resolution images.

5.3 Skip Convolution Cascade Module

Feature Visualisation In this subsection, the features of the oracle bone image output in different convolutional layers of the neural network are analyzed. The input oracle bone image“ rain ”is output through five stages of the VGG16 network, and each stage contains a different number of convolutional layers, namely 2, 2, 3, 3, 3. To have a more comprehensive view of the response of the features of the convolutional neural network in each convolutional layer to each region of the oracle bone image, the feature maps of the output of each stage of the VGG16 network are summed up. The feature maps output from each stage were accumulated. The accumulated feature maps are shown in Fig.4.

From Fig.4, the process of extracting the features of the oracle image by the convolutional neural network in different layers can be observed more clearly. In the shallow network, the feature maps outputted by the convolutional layer retain the low-level detail features of the image well and also have a strong response to the noise in the image. These low-level features are important for recognizing basic shapes and textures in the image. However, as the layers of the network deepen and the resolution of the image decreases, the textual details in the oracle image gradually become blurred in the feature map. This indicates that the feature maps output from the convolutional layers in the deep network discard a lot of textual detail information, and the extracted features are more abstract and complex. These advanced features focus more on the semantic information of the image rather than just the visual appearance of the image. It can be concluded that the low-level features output from the convolutional layer of the shallow network are more concerned with the details of the target in the image, while the high-level features output from the convolutional layer of the deep network are more concerned with the semantic information contained in the image.

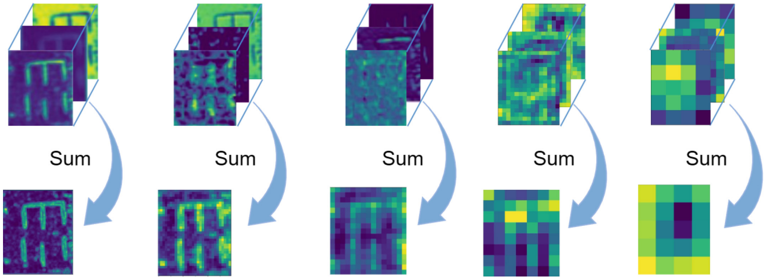


Fig. 4. The feature map output after the accumulation of convolutional layers in each stage of the VGG16 network

Design of Skip Connection and Convolutional Cascading Module In oracle bone inscription image retrieval, if only the low-level features of the image are utilized, the retrieval accuracy may not be guaranteed due to image changes and only the set of images with high similarity can be retrieved. If only high-level features of the image are utilized, the retrieval results will only take into account the semantic information of the oracle image and may ignore the specific details of the image and thus the accuracy of the retrieval. Therefore, to improve the performance of Oracle image retrieval, it is necessary to combine both low-level and high-level features to describe the images more comprehensively and to improve the accuracy and robustness of retrieval.

The feature extraction structure of ResNet consists of five stage layers, each of which contains a different number of convolutional and maximum pooling layers. As the number of network layers increases, the size of the extracted feature maps decreases. This process may lead to the disappearance of gradients, loss of feature information, and lack of local detailed features, which in turn affects the network's ability to distinguish between different oracle scripts. Especially when it is necessary to distinguish between samples with high similarity, the network may produce misclassification. In DenseNet [23], the connections between the input and output layers are much shorter close together, and each layer accepts the feature maps of all previous layers as inputs and also uses its own feature maps as inputs for all subsequent layers. This design effectively mitigates the gradient vanishing problem and enhances feature propagation and reuse. Inspired by DenseNet, this subsection introduces skip connections in the ResNet18 network and uses a "convolutional cascade module" to dimensionally splice the output of the previous stage with the output of the current stage. The module induces a path after each Stage layer and splices it dimensionally to form a cascade structure. This design induces features to flow and interact between different Stages, enhances feature transfer and utilization, and helps the network to better learn richer feature representations. The skip connection structure designed in this subsection is shown in Fig. 5.

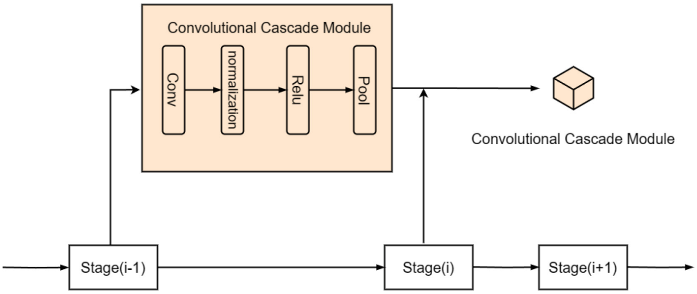


Fig. 5. The design of the skip connection structure and the composition of the convolutional cascade module

This design can be seen as introducing an additional "auxiliary path" in each ResNet block, which better preserves the local details of the features to increase the depth and complexity of the network, and helps to learn richer feature representations. As a result, the network is able to distinguish highly similar oracle characters more efficiently, thus improving the recognition rate. In this subsection, the proposed convolutional cascade module is to concatenate the output feature maps of the previous stage (i-1) with the output feature maps of the current stage (i) at the channel level, which is then used as the input of the next convolutional cascade module. This process requires that the summed feature maps are consistent in dimension, width, and height. The same operation is performed for each stage of the ResNet18 network, and the specific design is shown in Fig. 3.

6 Experimental Setup and Analysis

6.1 Experimental Environment and Experimental Configuration

The experiments in this paper were conducted under the Pytorch1.7.0 deep learning framework built-in Windows environment. The specific experimental environment: the CPU is Intel(R) Xeon(R) Platinum 8255C, the memory is 40G, the GPU is NVIDIA GeForce RTX 3080, the operating system is ubuntu18.04, and the programming language is Python 3.8. The experiments related to this paper's work use the same data format, i.e., adjust all the image sizes to 128*128*3. During the training process, the Batch Size is set to 32, the epoch is set to 30, and the Adam optimizer is used to update the model parameters, the size of the initial learning rate is set to 0.001, and the learning rate is periodically reduced according to the cosine function.

6.2 Evaluation Criteria

The evaluation metrics used in this paper are MP and MAP, where MP (Mean Precision) denotes the average value of accuracy, and a higher MP value indicates better retrieval. The calculation formula is shown below.

$$\mathrm{MP} = \frac{1}{\mathrm{C}} \sum_{1}^{\mathrm{c}} P@\mathrm{k} \tag{1}$$

where C denotes the number of search categories. P@k denotes the accuracy of the first k of each query.

$$\mathrm{P@k} = \frac{correct}{\mathrm{k}} \tag{2}$$

Where correct denotes the number of similar images retrieved correctly and k denotes the total number of images retrieved.

Mean Average Precision Mean MAP (mean average precision) [24]is the most commonly used evaluation metric in the field of image retrieval, which describes the ranking quality of the retrieval results.MAP represents the average precision (AP) of the query results accounted for by all the query-correct results summed up and divided by the total number of categories. The higher the value of MAP, the higher the retrieval results, and the higher the value of MAP. The MAP calculation formula is shown below.

$$\mathrm{MAP} = \frac{1}{\mathrm{C}} \sum_{1}^{c} AP@\mathrm{k} \tag{3}$$

where C denotes the number of retrieval categories, AP@k denotes the average precision of the first k per query

$$\mathrm{AP@k} = \frac{1}{\mathrm{N}} \sum_{1}^{\mathrm{N}} \frac{\mathrm{i}}{position(i)} \tag{4}$$

Where N denotes the number of retrieved images, i denotes the ith retrieved image, and position(i) denotes the position where the ith image is located.

6.3 Experimental Results and Analysis

In this section, the training set used is OBC219 and is divided into training and validation sets in the ratio of 4:1. The retrieval dataset used is TOBC40 divided from OBC306. The evaluation metrics are uniformly MP and MAP, and the highest scores in all experimental results are shown in bold, and the next highest scores are shown with underlined wavy lines. All scores are reported as percentages.

Replacement Convolution Experiment To verify the effectiveness of replacing the large convolution in the stage(0) layer of the ResNet network with a small convolution, this subsection conducts a comparison experiment on the TOBC40 retrieval dataset. In Table 1, the retrieval performance metrics (MP and MAP) for different model structures ResNet and ResNet-A (replacing 7*7 with three 3*3) are compared for multiple k values.

Table 1. Experimental results of Replacement convolutions on TOBC40

		MP			MAP	
model	k=5	k=10	k=25	k=10	k=50	k=100
ResNet	79.43%	73.31%	54.14%	86.45%	76.68%	68.82%
ResNet-A	**80.54%**	**74.12%**	**54.73%**	**87.13%**	**77.26%**	**69.37%**

From the experimental results in Table 1, it can be seen that among the two model structures, ResNet and ResNet-A, ResNet-A achieves slightly better performance in MP and MAP values at all k values, with the most significant improvement in MP at k=5 and k=10. At k=5 and k=10, the performance improvement of ResNet-A is more obvious, and the improvement of MP is more significant, which is about 1.11% and 0.81% respectively. This indicates that in the retrieval experimental results of the TOBC40 dataset, the number of retrieved correct images is more in the results obtained by the ResNet-A model structure.

Experiments on Skip Connections and Convolutional cascade modules Finally, to evaluate the effectiveness of the method of adding skip convolutional cascade modules to the ResNet network, experimental comparisons of the improved network with the ResNet18 and DenseNet121 networks are conducted in this paper. The experimental results are shown in Table 2.

Table 2. Verification of the validity of the skip convolution cascade module on TOBC40

		MP			MAP	
model	k=5	k=10	k=25	k=10	k=50	k=100
ResNet	79.43%	73.31%	54.14%	86.45%	76.68%	68.82%
Improved ResNet	80.79%	74.54%	55.03%	89.12%	77.56%	69.84%
DenseNet121	**80.85%**	**74.63%**	**56.34%**	**89.24%**	**77.76%**	**69.91%**

Table 2 shows the performance comparison between the pre-and post-improved ResNet networks, as well as the DenseNet121 network on the TOBC40 dataset with different numbers of images (k), returned. As can be seen from the table, the improved ResNet network outperforms the pre-improved ResNet network for all k values. For the DenseNet121 network, its performance is comparable to the improved ResNet network for most of the keypoint values, but slightly better than the improved ResNet network at k=5 and k=10. It can be concluded that the addition of the jump-convolution cascade module significantly improves the performance of the ResNet network on the two different datasets, especially on the TOBC40 dataset. The improved ResNet network shows a stable performance improvement for most k-values, which indicates that the jumping convolutional cascade module is an effective way to enhance the performance of the ResNet network.

In order to more intuitively feel the difference between the feature maps before and after adding jump connections, the output feature maps of each feature extraction layer before and after the improvement were summed up, and then the summed-up feature maps were compared. The accumulated feature maps are shown in Fig.6, from which the richness in details of the improved feature maps can be clearly observed.

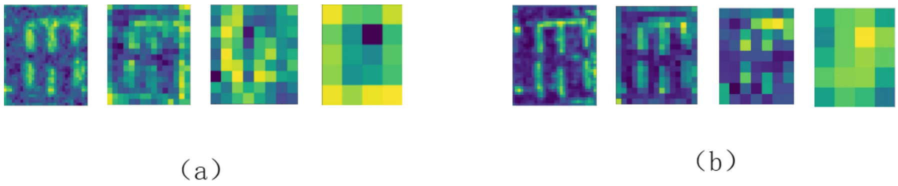


Fig. 6. The result of the accumulation of the output feature map after each feature extraction layer before and after the improvement, where (a) and (b) represent the ResNet network before and after the improvement, respectively.

From the figure, it can be clearly observed that the improved output cumulative feature picture has higher clarity than the pre-improved one. This indicates that the improved network is able to better fuse the shallow local detail information with the deep global semantic information, which makes the feature image richer in content and improves the effective utilization of the features. This improvement helps to enhance the ability of the network to extract and represent image features and also improves the accuracy and robustness of the network in distinguishing and recognizing images.

Comparison Experiment Based on the ResNet network, this paper improves the ResNet network in two aspects and confirms the effectiveness of the proposed method through comparative experiments and visual analysis. In this section, the first two improved modules will be integrated into the ResNet18 network and the improved network will be called OBINet. In this section, we will discuss the experimental results of the OBINet network compared to the existing network used for Oracle image feature extraction. In this experiment, other conditions are fixed to be consistent to verify the influence of different feature extraction networks on the experiment.

Table 3. Comparison of OBINet and other network models on TOBC40

		MP			MAP	
model	k=5	k=10	k=25	k=10	k=50	k=100
VGG16[2]	77.15%	71.49%	52.34%	84.24%	73.03%	66.54%
ResNet18[2]	879.43%	73.31%	54.14%	86.45%	76.68%	68.82%
inception-v4[25]	79.97%	73.54%	55.79%	87.53%	76.87%	68.94%
ResNeSt[18]	80.25%	74.32%	57.78%	88.38%	77.12%	69.47%
Yao[8]	82.14%	75.37%	58.47%	89.15%	78.24%	71.18%
OBINet(Ours)	**83.26%**	**76.53%**	**59.12%**	**90.68%**	**79.29%**	**72.47%**

Table 3 shows the retrieval performance of the OBINet model compared to several other network models on the TOBC40 dataset. The data in the table gives the retrieval results of different models for different numbers of returned images (k=5, k=10, k=25,k=50, k=100) in the form of MP, and MAP. From the table, it can be seen that the OBINet model outperforms the other models in all the evaluated metrics, especially at the larger number of retrieval k=100, the OBINet model has the highest MAP value of 74.47%, which is 3.65% higher than the MAP value of the pre-improvement network with ResNet18 as the feature extraction network. This indicates that the OBINet model has a stronger retrieval ability on the TOBC40 dataset.

7 Summary and Outlook

Image retrieval of oracle bone inscription is of great significance. It not only helps to protect and study the oracle bone heritage, but also improves the efficiency and level of oracle bone research, and promotes the process of digitization and intelligence of oracle bones. To retrieve images similar to the query image more accurately, this study firstly adopts a Siamese network-based image retrieval method for oracle bone inscription, which learns feature representations that distinguish between similar and dissimilar images and achieves efficient retrieval. Next, the existing dataset is divided for the problem of unbalanced sample distribution, which provides a practically usable retrieval dataset for the field of oracle bone inscription image retrieval. Finally, a backbone network for oracle bone inscription image retrieval is proposed, which is optimized on the basis of the ResNet network: the initial 7*7 large convolutional layers in the ResNet network are replaced by three 3*3 small convolutional layers; a skip convolutional cascade modular structure is designed to transfer and fuse features from the shallow layer of the network to the deeper layer of the network;

At present, the research on oracle bone inscription image retrieval is still in the preliminary stage, and there are certain challenges and limitations. For example, the low resolution of oracle bone inscription images and the severe image background noise all have some impact on the accuracy of image retrieval. Therefore, future research can further explore more efficient and accurate image retrieval algorithms to improve the effectiveness of image retrieval of oracle bone inscription.

Acknowledgements.. This study was Supported by the National Natural Science Foundation of China(NO.62266044,62061045). It was also supported by the "Tianshan Talents" Leading Talents Program for Scientific and Technological Innovation in Xinjiang Uygur Autonomous Region (2023TSYCLJ0025), and the Open Project of Key Laboratory of Oracle Bone Inscription Information Processing, Ministry of Education (OIP2021E004).

References

1. H. Zhan and Y. Qi, Chinese character image retrieval based on moment invariants and shape context. 2015 IEEE International Conference on Computer and Communications (ICCC), Chengdu, China, 2015, pp. 146-150
2. S. Huang, H. Wang, Y. Liu, X. Shi, and L. Jin, OBC306: A Large-Scale Oracle Bone Character Recognition Dataset. 2019 International Conference on Document Analysis and Recognition (ICDAR), Sydney, NSW, Australia, 2019, pp. 681-688
3. Q. Zhang, Z. Wang, X. Hu, and R. Chen, A Content-Based Image Retrieval Scheme for Encrypted Domain Using Feature Fusion Deep Supervised Hash. 2023 IEEE International Conference on Sensors, Electronics and Computer Engineering (ICSECE), Jinzhou, China, 2023, pp. 34-39
4. Jing, X., Gao Feng, W., Qinxia,: Research on Semantic Mining for Large-scale Oracle Bone Inscriptions Foundation Data. New Technology of Library and Information Service **31**(2), 7–14 (2015)
5. T. Lin, Method of oracle bone inscription image retrieval based on Siamese neural network(in Chinese). Xiamen University, 2020
6. Liu, G., Wang, Y.: Oracle character image retrieval by combining deep neural networks and clustering technology. Int. J. Comput. Sci. **2**, 199–206 (2015)
7. Han X, Bai Y, Qiu K, et al, IsOBS: An information system for oracle bone script. Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demon- stations. Association for Computational Linguistics, 2020: 227-233
8. Zhixi.Yao, Research on Oracle Bone Script Image Recognition and Retrieval Based on Multi-Strategy Enhancement(in Chinese). Xinjiang University, 2022
9. K. R. N. Aswini, S. P. Prakash, G. Ravindran, T. Jagadesh and A. V. Naik, An Extended Canberra Similarity Measure Method for Content-Based Image Retrieval. 2023 International Conference on Evolutionary Algorithms and Soft Computing Techniques (EASCT), Bengaluru, India, 2023, pp. 1-5
10. Kumar, G.V.R.M., Madhavi, D.: Stacked Siamese Neural Network (SSiNN) on Neural Codes for Content-Based Image Retrieval. IEEE Access **11**, 77452–77463 (2023)

11. Sumbul, G., Ravanbakhsh, M., Demir, B., Relevant, A., Hard and Diverse Triplet Sampling Method for Multi-Label Remote Sensing Image Retrieval.: IEEE Mediterranean and Middle-East Geoscience and Remote Sensing Symposium (M2GARSS). Istanbul, Turkey **2022**, 5–8 (2022)
12. N. Carlevaris-Bianco and R. M. Eustice, Learning visual feature descriptors for dynamic lighting conditions. 2014 IEEE/RSJ International Conference on Intelligent Robots and Systems, Chicago, IL, USA, 2014, pp. 2769-2776
13. Z. Pan, X. Bao, Y. Zhang, B. Wang, Q. An, and B. Lei, Siamese Network-Based Metric Learning for SAR Target Classification. IGARSS 2019 - 2019 IEEE International Geoscience and Remote Sensing Symposium, Yokohama, Japan, 2019, pp. 1342-1345
14. F. Radenović, G. Tolias, and O. Chum, Fine-Tuning CNN Image Retrieval with No Human Annotation. in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 41, no. 7, pp. 1655-1668, 1 July 2019
15. Razavian A S, Sullivan J, Maki A, et al, A Baseline for Visual Instance Retrieval with Deep Convolutional Networks.ITE Transactions on Media Technology and Applications, 2014, 4(3)
16. Chicco, D. (2021). Siamese Neural Networks: An Overview. In: Cartwright, H. (eds) Artificial Neural Networks. Methods in Molecular Biology, vol 2190. Humana, New York, NY
17. Linchang Zhao, Zhaowei Shang, et al, Siamese networks with an online reweighted example for imbalanced data learning, Pattern Recognition, Volume 132,2022
18. Yafei MAO, BI Xiaojun. Rubbing oracle bone character recognition based on improved ResNeSt network. CAAI Transactions on Intelligent Systems, 2023, 18(3): 450-458
19. Zhang, Y.-K., Zhang, H., Liu, Y.-G., et al.: Oracle character recognition based on cross-modal deep metric learning. Acta Automatica Sinica **47**(4), 791–800 (2021)
20. K. He, X. Zhang, S. Ren and J. Sun, Deep Residual Learning for Image Recognition. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 2016, pp. 770-778
21. C. Szegedy, V. Vanhoucke, S. Ioffe, et al, Rethinking the Inception Architecture for Computer Vision. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 2016, pp. 2818-2826
22. Shen, Y. et al. (2020). Enabling Deep Residual Networks for Weakly Supervised Object Detection. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, JM. (eds) Computer Vision – ECCV 2020. ECCV 2020. Lecture Notes in Computer Science, vol 12353. Springer, Cham
23. Huang, G., Liu, Z.: Maaten L D, et al, Densely connected convolutional networks. IEEE Conference on Computer Vision and Pattern Recognition (CVPR). **2017**, 2261–2269 (2017)
24. Christopher D, Manning P R, Schutze H. Introduction to information retrieval. Cambridge University Press, 2008
25. Harbin Wang, Research on Oracle Bone Script Detection and Recognition Based on Deep Learning(in Chinese). South China University of Technology, 2019

Novel Clustering Aggregation and Multi-grained Alignment for Image-Text Matching

Shuming Zhang, Xiao-jun Wu(✉), Tianyang Xu, and Donglin Zhang

School of Artificial Intelligence and Computer Science, Jiangnan University, Wuxi, China
6213113146@stu.jiangnan.edu.cn,
{wu_xiaojun,tianyang.xu,zhangdlin}@jiangnan.edu.cn

Abstract. As a challenging multi-modal task, image-text matching continues to be an attractive topic of research. The essence of this task lies in narrowing down the semantic disparity between vision and language to align them better. Existing works have either focused on coarse-grained alignment between global images and texts or fine-grained alignment between salient regions and words. However, they do not distinguish between considering related and redundant pairs (i.e., regions with no matching words or pairs with low relevance). We thereby propose a Novel Clustering Aggregation and multi-grained Alignment network (NCAA), which utilizes cross-modal contextual clustering to group regions based on semantic information consistent with the text content. Specifically, we leverage textual fragments as clustering centers, similarity between regions and fragments as propagation medium and delicately devise two mask mechanisms for simultaneous and distinguishable consideration of both related and redundant pairs. Two alignment modules of different granularities are also introduced to achieve the multi-grained alignment. By incorporating both global and local similarity into the training and inference phases, our model attains further enhancements. Finally, we conduct extensive experiments on two benchmark datasets, Flickr30K and MSCOCO, demonstrating the efficacy of our framework.

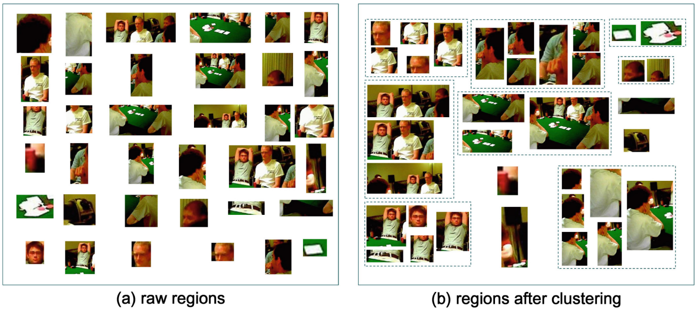
Keywords: Image-Text Retrieval · Multi-modality · cross-modal contextual clustering

1 Introduction

As a fundamental multimodal task, image-text matching primarily focuses on reasoning about image and text features to narrow the semantic gap between these two modalities. The task is closely related to various multimodal tasks, such as image captioning [15] and Artificial Intelligence Generated Content (AIGC) [7] task. Recently, image-text matching has gained more and more attention and many related works have been proposed. However, the persistent semantic disparities between visual content and textual descriptions remain a significant challenge. We categorize previous methods designed to address this problem from the perspective of alignment granularity into three classes: coarse-grained alignment, fine-grained alignment, and multi-grained alignment.

Coarse-grained alignment methods [14,21,24] aim at discovering semantic correspondences of the entire sentences and images, representing the global-level alignment.

A. Antonacopoulos et al. (Eds.): ICPR 2024, LNCS 15321, pp. 63–79, 2025.
https://doi.org/10.1007/978-3-031-78305-0_5



(a) raw regions

(b) regions after clustering

Fig. 1. Illustration about our cross-modal contextual clustering module, revealing that this method can not only cluster regions containing the same semantic information together, but also separate unrelated regions into individual groups.

The prevalent framework for coarse-grained alignment methods typically involves a dual-branch deep neural network, with one branch dedicated to images and another to text. Due to the independence of these two branches, inference can be conducted separately for each branch when new data arrives. This design allows for fast inference speeds but may not effectively bridge the semantic gap between the two modalities, especially for complex image-text pairs.

Fine-grained alignment methods [5,13,25] focus primarily on aligning salient region-word pairs, which stand for the local-level alignment. Most of them are based on attention mechanisms. Here, cross-attention mechanisms are applied for fusing or aligning features between different modalities. The inter-modal interactions facilitated by cross-attention mechanisms contribute to significant performance improvement. However, feature interactions between modalities can be highly computationally intensive. As a consequence, fine-grained alignment methods may exhibit slower performance than coarse-grained alignment methods during the inference.

Coarse- and fine-grained alignment methods have their own advantages and disadvantages when used separately. Therefore, multi-grained alignment methods [16,26] combining advantages of both coarse- and fine-grained considerations have become prominent, which makes these two grained alignment methods complement each other, not only enhancing retrieval speed but also striking a balance between performance and efficiency.

However, despite these extensive advancements, designing interaction patterns within each modality and between different modalities have not been completely conquered. Discovering optimal solutions to bridge the complex cross-modal semantic gaps remains an ongoing challenge. To better address these issues, we introduce the Novel Clustering Aggregation and multi-grained Alignment network (NCAA) which falls within the multi-grained alignment category. Unlike previous mainstream approaches that primarily rely on attention mechanisms, based on our investigations, we are the first to utilize cross-model clustering algorithm in image-text matching. As shown in Figure 1 (a), many existing methods do not consider semantic associations of visual region features, resulting in a cluster of these features that lack interconnections. Furthermore, some regions exhibit low relevance to word features, making them appear redundant in the alignment. Therefore, we employ a novel clustering approach, including two different propagation strategies to cope with related and redundant regions, respectively. Related regions can be assigned to the same semantic clusters while redundant regions are also categorized individually, as illustrated in Figure 1(b). It is worth mentioning that we have also introduced the multi-glance reasoning module for fine-grained alignment and the global guidance module to accomplish coarse-grained alignment. Both local and global perspectives are considered in the training and inference phases. Particularly, we employ a two-stage inference approach to leverage local and global similarities more comprehensively. Compared to previous methods, our NCAA outperforms them and achieves state-of-art performance. Detailed ablation experiments and visualizations confirm the effectiveness of our model.

In summary, our contributions can be summarised as follows:

- We propose a novel cross-model contextual clustering by ingeniously utilizing textual fragments as clustering centers and two aggregation strategies to cluster related and redundant regions respectively.
- We intricately design two alignment modules with different granularities to implement multi-grained alignment.
- Experimental results show that our method NCAA achieves SOTA performance on several public benchmarks.

2 Related Work

2.1 Multi-grained Alignment

Many early studies are primarily based on the coarse-grained alignment strategy, which typically calculates the global similarity between image and text. These approaches frequently enhance performances through the exploration of various training loss functions. Among them, VSE++ [6] is the first to consider the hard negative sample mining and the enhanced Max Hinge Loss. Some researches [4,9,10] also place emphasis on aggregation strategies for obtaining global features with richer semantics.

With the development of attention mechanisms, many works have shifted towards fine-grained alignment strategy recently. SCAN [8] , which is a very classic work, utilizes differently weighted regions and words to infer the local similarity. Subsequently, a plethora of work [19,20,27] has been inspired by it. For instance, [20] contemplate the importance of different locations within each region. [27] explicitly considers both positive and negative matching segments to jointly measure the similarity. Thus, it can be seen that fine-grained alignment methods are dedicated to designing more sophisticated models.

Coarse-grained alignment methods offer fast inference speed but lack high precision, while fine-grained alignment methods exhibit the opposite characteristics. Therefore, multi-grained alignment methods that combine these two alignment methods have become prominent. DIME [16] ,which is our baseline, designs a routing mechanism to realize dynamic modality interaction, ultimately providing a comprehensive consideration of global and local alignment. [18] develops a Scene Concept Graph as scene common-sense knowledge to achieve multi-grained alignment. Our model falls under the category of multi-grained alignment methods, attaining an excellent balance between performance and efficiency.

2.2 Contextual Clustering

Nowadays, mainstream deep learning frameworks predominantly rely on convolution or attention mechanisms. Clustering algorithms are no longer commonly employed in the field of deep learning anymore. However, researches continue on efficient clustering algorithms [1,12,22]. SuperPixel [22] directly learns superpixel segmentation results from input images in an end-to-end manner. SLIC [1] converges faster, restricting clustering operations within local regions and initializes K-means centers uniformly. Inspired by [12], combining cross-model contextual clustering with image-text matching is related to our work. Nonetheless, to the best of our knowledge, our work represents the first attempt to apply the cross-model clustering algorithm to image-text matching.

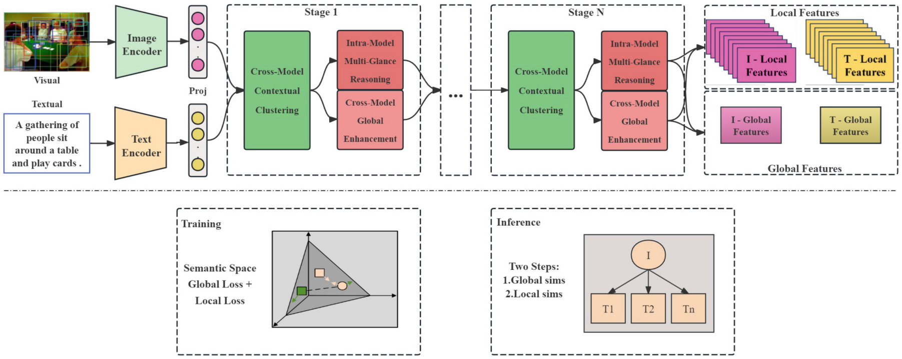


Fig. 2. Architecture of our proposed NCAA model. Input images and texts are initially encoded and subsequently mapped into a shared embedding space. Three modules are devised in blocks with different colors and stacked in a hybrid connected way to form a multi-grained alignment interaction stage. During training, the global loss and local loss simultaneously optimize the network. During inference, a two-stage method based on global and local similarity achieves both accuracy and efficiency.

3 Methodology

In this chapter, we extensively discuss each component of our model, as depicted in Figure 2. Initially, in Section 3.1, we delve into the representation learning of two modalities. Subsequently, in Section 3.2, we introduce three alignment modules, comprising the context clustering module for pre-alignment preparation, the multi-glance reasoning module for local alignment, and the global guidance module for global alignment. Finally, in Section 3.3, we present the objective function and the two-stage inference strategy that we employ.

3.1 Representation Learning

Visual Representation Given an input image *I*, following [2], our visual representation learning is established upon the Faster R-CNN detector [17]. We select K salient image proposals, where the local region feature vector $r_i \in \mathbb{R}^d$ along with the bounding box vectors $p_i \in \mathbb{R}^4$ for each proposal. The aspect ratio and area are additionally incorporated into bounding box vectors to enhance the position information. We then concatenate r_i and p_i and feed them into a fully connected layer (FC) to obtain region-level visual features denoted as $v_i \in \mathbb{R}^D$. Simply computing the average over all local region features as the global image feature tends to lose a lot of fine-grained details. Therefore, we employ the average and region features as the query and key of the self-attention mechanism to obtain the global image feature v_0 by aggregating local information. Finally, we define the visual representation as $I = \{v_0, v_1, ..., v_K\}$.

Textual Representation Currently, the best-performing language model is based on the Transformer architecture. Given a sentence *T* containing L words, the pre-trained Bert model [3] is utilized to extract word-level textual features which can be denoted as $T = \{e_1, e_2, ..., e_L\}$, where $e_i \in \mathbb{R}^{768}$ and 768 is the hidden-size of Bert. Afterward, we deploy a FC to transform e_i to a *D*-dimensional vector indicated as $\tilde{t}_i$. Just like humans understand a sentence, phrases can help us better grasp the semantic information of a sentence. In practical terms, this requires considering several words at the same time. Consequently, we exploit series convolutions with different kernels to acquire phrase-level features that are context-enhanced. Then, we concatenate the features generated by different kernels and feed them into another FC layer to map them into the D-dimensional space as $\hat{t}_i$. We sum the word-level and phrase-level features to acquire the local textual features as $t_i = \hat{t}_i \oplus \tilde{t}_i$. Finally, we employ the same strategy as the visual branch to obtain sentence-level global textual features denoted as t_0. So the whole textual representation can be defined as $T = \{t_0, t_1, ..., t_L\}$.

3.2 Alignment Modules

Our proposed NCAA is multi-grained alignment method. We deliberately refrain from employing complex network structures to ensure efficient model inference. Instead, we design three alignment modules in a simple model, which we describe in details below.

Cross-Model Contextual Clustering (CCC) Given image-text pairs (I, T), it contains a wealth of related or redundant pairs. The objective of the cross-modal contextual clustering module is to group these two types of pairs and process them separately using corresponding strategies. The cross-modal contextual clustering module represents the core contribution of our work, as depicted in Figure 3.

Holistically, this module is designed specifically for local features. The operation involves grouping region features into clusters, where semantically similar region features are aggregated and mutually propagated. We implement it based on the local similarity between regions and textual fragments and two allocation strategies to assign regions with semantic relevance and redundancy to different clusters. We first linearly project v_i to v_i^s for similarity computation and then calculate the contextual weight on

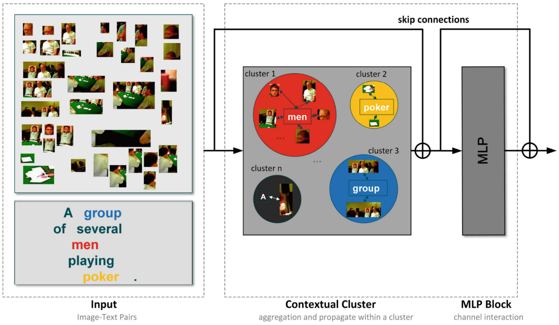


Fig. 3. Illustration of our proposed Cross-model Contextual clustering module.

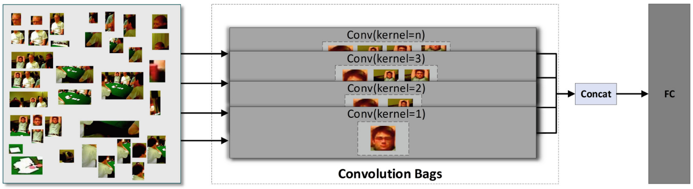


Fig. 4. Illustration of our proposed Intra-model Multi-Glance reasoning module.

each region with respect to each word as follows:

$$\alpha_{ij} = \frac{\exp\left(\lambda \hat{s}_{ij}\right)}{\sum_{i=1}^{K} \exp\left(\lambda \hat{s}_{ij}\right)}, \tag{1}$$

here the weight α_{ij} is calculated by the softmax function with a temperature parameter λ. $\hat{s}_{ij} = \left[s_{ij}\right]_{+} / \sqrt{\sum_{j=1}^{L} \left[s_{ij}\right]_{+}^{2}}$ intents to normalize the similarity s_{ij} between v_i^s and t_j. $[x]_{+} = max\,(x, 0)$.

Then, we can obtain the attended features that are employed as the cluster center $c^s \in \mathbb{R}^{L \times D}$ by: $c_j^s = \sum_{i=1}^{K} a_{ij} v_i^s$ and the local pair-wise cosine similarity matrix $S \in \mathbb{R}^{L \times K}$ between v_i^s and the center feature c_j^s. Subsequently, we assign regions to centers, resulting in L clusters. It is worth noting that each cluster may contain a different number of regions. During feature aggregation, our work can simultaneously focus on both related and redundant regions.

The strategy of distinguishing relevant and redundant regions before feature propagation is performed through two masks. The magnitude of similarity S reflects the degree of matching between local features. For regions that contain semantically similar information and match words, we aim to group them together as much as possible. Meanwhile, for redundant regions that do not match clustering centers, it is preferable to minimize the consideration of redundant information, thus reducing its negative effect on alignment. So we design two mask mechanisms to mine the intrinsic connection of regions and the semantic connections between regions and words as follows:

$$Mask(S) = Mask_m([\|S\|_1 - \mu]_+) + Mask_r(Max(\|S\|_1 - \mu)), \tag{2}$$

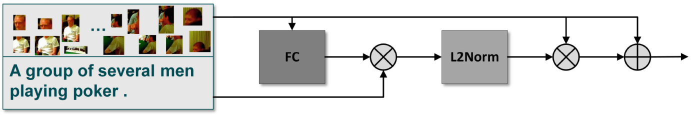


Fig. 5. Illustration of our proposed Cross-model Global Enhancement module.

where $\|\cdot\|_1$ means L1 normalization, and μ is an empirically manually adjusted relevance boundary. If the similarity is greater than this threshold, it is considered as matching pairs; otherwise, as redundant pairs. $Mask_m(\cdot)$ is the matched mask that when the input outweighs 0, it equals itself, otherwise it is 0. Meanwhile, $Mask_r(\cdot)$ is the redundant mask that we only select the maximum value belonging to the mismatched ones so that its value is 1; otherwise it is 0. For redundant regions, we aim to consider them separately, so we select the maximum value in the redundant pairing that is less than the threshold value to participate in the feature aggregation. This can reduce the negative effect of redundant regions on image text retrieval, but at the same time preserve their representation ability to enrich the global features of the image and improve the image understanding ability. Integrating these two masks to obtain the final $Mask(\cdot)$, where the original 1s remain unchanged, and the original 0s are replaced with $-\infty$ for the next activation function.

We dynamically aggregate regions to their respective text centers using masked similarity $\tilde{S} = Mask(S)$. We also map the region features v_i to the value space to get v_i^v and obtain a value center c^v using the same calculation method as for c^s based on the v_i^v. The similarity between i-th region and the cluster center is represented as $\tilde{s}_i$ (a subset in $\tilde{S}$). The aggregated feature f is given by :

$$w_i = \mathrm{Softmax}\left(\alpha \tilde{s}_i + \beta\right), \tag{3}$$

$$f_i = \frac{1}{C}\left(c^v + \sum_{i=1}^{K} w_i * v_i^v\right), \quad \text{s.t.,} \quad C = 1 + \sum_{i=1}^{K} w_i, \tag{4}$$

here α and β are learnable scalars to scale and shift the similarity. To control the magnitude, the aggregated feature f_i is normalized by a factor of C. We incorporate the value centers c^v, both for extracting global information and accelerating neural network fitting to maintain training stability. The aggregated feature f_i is then adaptively propagated among each region in a cluster based on the similarity weight w_i. By doing so, regions can interact with all other regions in the same cluster, propagating features to each other and enabling a better reasoning of the the aggregated features. We update it by a FC and skip connection, which is formulated as :

$$v_i' = v_i + \mathrm{FC}\left(w_i * f_i\right), \tag{5}$$

Finally, we feed the propagated features v_i' into an affine transformation composed of MLP to facilitate channel-wise communications. Likewise, a shortcut connection is employed here to obtain the last region features set I_{cc}.

Intra-Model Multi-Glance Reasoning (IMR) Just as humans may focus on one or multiple regions in a single glance while looking at an image. For better reasoning region features, we consider the interactions of regions from different views as shown in Figure 4. This allows each region to capture semantic information from other regions in different combinations thus grasping the semantic dependencies. Given the clustered region features I_{cc}, a parallel set of 2D convolutions is employed to calculate the glance-specific vectors for each region. The convolution kernels are used to symbolize the receptive field. For instance, if the size of the kernel is 3, it means that 3 regions are simultaneously weighted and summed. In order to balance the performance of the model and the computational cost, according to our experience, in this module, we choose the number of convolution kernels as [1,2,3]. Then, we concatenate the feature maps of these various convolutions and pass the result through a FC layer to obtain the glance-specific region features. Concretely, we capture the intra-modal dependencies from different subspaces as,

$$I_{mr} = FC(Concat[I'^{1}_{cc}, I'^{2}_{cc}, ..., I'^{n}_{cc}]), \tag{6}$$

where $Concat(\cdot)$ represents the concatenation operation across the feature dimension. I'^{n}_{cc} means the outputs of the n-th convolution.

Cross-Model Global Enhancement (CGE) Although the above two proposed modules align local features from the fine-grained perspective, providing abundant clues, it is still necessary to adopt the global features which are consolidated contextual information and high-level semantics as guidance to refine and enhance the semantic level of the local features as the Figure 5 illustrates and as follows:

$$I'_{ge} = FC(I_{cc}) * t_0, \tag{7}$$

$$I_{ge} = (1 + \left\|I'_{ge}\right\|_2) * I_{cc}, \tag{8}$$

where t_0 represents the global textual feature. $\|\cdot\|_2$ denotes L2-Normalization across the dimension of features.

Multi-grained Matching Combining the three modules elaborated above in a hybrid-modal interaction forms one stage. Our proposed IMR module and CGE module align image and text features from local and global perspectives respectively, narrowing the semantic gap between two modalities. The outputs of these two modules from the previous stage are averaged as the visual input for the next stage. After a series of iterations, the outputs $(\hat{I}_{ge}, \hat{I}_{mr})$ in the N-th stage serves as the enhanced local image feature, capturing a more comprehensive interaction for similarity prediction.

For coarse-grained global matching, we first apply the same function that we use to obtain v_0 on the average of $\hat{I}_{ge}$ and $\hat{I}_{mr}$ to acquire the enhanced image features $\hat{I}^g$. Adding it and the original image feature v_0 accompanied by a batch normalization obtains the final global image feature $\tilde{v}_0$. For any two samples, the global coarse-grained similarity is defined as :

$$S_g(I, T) = \frac{\tilde{v}_0^T t_0}{\left\|\tilde{v}_0^T\right\|_2 \|t_0\|_2}. \tag{9}$$

For fine-grained local matching, the final local feature is accompanied by summing all local features together by $\hat{I} = BN(I + \hat{I}_{ge} + \hat{I}_{mr})$. Let $\tilde{I} = \{\tilde{v}_1, ..., \tilde{v}_K\}$ and

$T = \{t_1, ..., t_L\}$ be the local image and text features respectively. we first compute the semantic relevance scores as:

$$r_{ij} = \frac{t_i \hat{v}_j^{\mathrm{T}}}{\|t_i\|_2 \, \|\hat{v}_j\|_2}, i \in [1, L], j \in [1, K], \tag{10}$$

here r_{ij} reflects the cosine similarity between textual fragment t_i and visual region $\hat{v}_j$. Afterward, the local similarity is computed by decomposing the relevance matrix r_{ij} through the max-sum pooling:

$$S_l(I, T) = \sum_{i=0}^{L} \max_{j \in K} r_{ij}. \tag{11}$$

In the max-sum pooling function, the max operation computes the max over the regions intended for finding the most matching visual region for each textual fragment. Simultaneously, the summation operation is utilized to add the best matching similarities of all textual fragments as the local similarity between two modalities.

3.3 Objective Function

Mixed Training The objective function practiced in this paper is consistent with the approach adopted in many previous works [27], employing multi-modal contrastive learning method with the hinge-based bidirectional triplet ranking loss for end-to-end training. The specific process is as follows:

$$\begin{aligned} \mathrm{L}(I, T) = & [\delta - S(I, T) + S(I, T_{-})]_{+} \\ & + [\delta - S(I, T) + S(I_{-}, T)]_{+}, \end{aligned} \tag{12}$$

where δ is a margin hyperparameter which forces the model to strive to make the distance value between anchor I and negative example T_{-} larger while making the distance value between anchor I and positive example T smaller. Conversely, this holds true when the anchor is T. $S(\cdot)$ denotes the similarity function. The corresponding hardest negative text is $T_{-} = argmax_{j \neq T} S(I, j)$ and the hardest negative image is $I_{-} = argmax_{i \neq I} S(i, T)$ in a mini-batch .

Global matching is efficient but ignores local fine-grained details. Local matching achieves high accuracy but lacks global high-level semantics. However, applying either of them independently cannot strike a balance between efficiency and performance. Our model mixes these two autonomous approaches, achieving multi-grained matching. Given the global coarse-grained image-sentence similarity $S_g(I, T)$ and local fine-grained region-fragment similarity $S_l(I, T)$, the overall training loss is as follows:

$$\mathrm{L} = \mathrm{L}_g(I, T) + \mathrm{L}_l(I, T), \tag{13}$$

here $S_g(I, T)$ and $S_l(I, T)$ serve as the similarity functions of $\mathrm{L}_g(I, T)$ and $\mathrm{L}_l(I, T)$ respectively.

Two-stage Inference To balance the trade-off between efficiency and accuracy, we adopt a two-stage inference strategy. Simply put, first, the global similarity, including the entire dataset is ranked and sorted, which can be done efficiently. Then the top m results are re-ranked by the local similarity. With these two-step strategies, we can leverage the advantages of these two matching mechanisms in the inference stage to quickly and accurately retrieve results based on queries. At last, the mixed similarity composed by the global and local matching is applied as the final similarity.

More professionally, given a text query t and an image dataset containing M images, we first obtain a subset of the top m images (where $m \ll M$) with the highest scores through global matching. Then, we retrieve the final most relevant image through reranking by local matching. The specific process is as follows:

$$\arg\max_{x \in m} S_l(I,T) + S_g(I,T). \tag{14}$$

4 Experiments

In this chapter, we display the experiment results of our model NCAA on two benchmark databases to verify the superiority of our model. We conduct standard experiments in two scenarios, including I2T (retrieving the most relevant sentence given an image) and T2I (matching the most semantically consistent image with a given text query). Additionally, a substantial number of comprehensive ablation experiments and visualization results are also presented in this section.

Table 1. Performance comparison between our proposed NCAA and several modals on the Flickr30K and MS-COCO datasets. The best performance is highlighted in bold.

Methods	Flickr30K							MSCOCO(1K)						
	Text Retrieval			Image Retrieval			R@sum	Text Retrieval			Image Retrieval			R@sum
	R@1	R@5	R@10	R@1	R@5	R@10		R@1	R@5	R@10	R@1	R@5	R@10	
PFAN [20]	70.0	91.8	95.0	50.4	78.7	86.1	472.0	76.5	96.3	**99.0**	61.6	89.6	95.2	518.2
VSRN [10]	71.3	90.6	96.0	54.7	81.8	88.2	482.6	76.2	94.8	98.2	62.8	89.7	95.1	516.8
SGRAF [5]	77.8	94.1	97.4	58.5	83.0	88.8	499.6	79.6	96.2	98.5	63.2	90.7	96.1	524.3
CMCAN [26]	79.5	95.6	97.6	60.9	84.3	89.9	507.8	81.2	96.8	98.7	65.4	91.0	96.2	529.3
TERAN [13]	79.2	94.4	96.8	63.1	87.3	92.6	513.4	77.7	95.9	98.6	65.0	91.2	96.4	524.8
DIME [16]	81.0	95.9	98.4	63.6	88.1	93.0	520.0	78.8	96.3	98.7	64.8	91.5	96.5	526.6
NAAF [27]	81.9	**96.1**	98.3	61.0	85.3	90.6	513.2	80.5	96.5	98.8	64.1	90.7	96.5	527.2
RCAR [4]	82.3	96.0	98.4	62.6	85.8	91.1	516.2	80.9	**96.9**	98.9	**65.7**	91.4	96.4	530.2
NCAA	81.4	95.3	98.2	62.6	87.5	92.8	517.8	80.5	96.1	98.2	64.4	90.9	96.1	526.2
NCAA(ensemble)	**83.1**	96.0	**98.8**	**64.1**	**88.5**	**94.0**	**524.5**	**81.4**	**96.9**	**99.0**	65.4	**91.7**	**96.7**	**531.1**

4.1 Datasets

Flickr30K [23]: Flickr30K is a dataset curated from Flickr, comprising 31,783 images and each image has five corresponding descriptive sentences. Following [8], we split the dataset into 1000 test images, 1000 validation images and the rest training images.

MSCOCO [11]: The MSCOCO dataset is an important large-scale computer vision dataset that is equally substantial for vision-language tasks. This dataset contains 123,287 images, each with 5 annotated sentences. We split the dataset into 5000 images for testing, 5000 images for validation and the rest for training. We report the evaluation results directly computed on the full 5K test images (**COCO-5K**) and the matching result by an average over 5 independent folds, each composed of 1000 test images (**COCO-1K**).

4.2 Protocols

The rank at Top-K (R@K) is an evaluation metric widely used in the information retrieval domain. R@K indicates the proportion of ground truth contained in the top K samples of the retrieval results. A higher R@K value indicates a more accurate match. R@1, R@5, R@10 are used to quantitatively test the performance of our model. Additionally, R@sum, obtained by summing up all the metrics, is used for comparison with other competitive works.

4.3 Implementation Details

As most, we extract 36 region proposals for each image and 32 words for each sentence. The visual and textual features are mapped into a shared 256-dimensional embedding space. For the manually set hyperparameters, the temperature parameter λ in Eqn. (1) is set to 9. The margin parameter δ in Eqn. (12) is set to 0.2. In addition, the boundary parameter μ in Eqn. (2), representing the distinction between related and redundant regions, is set to 0.2 based on previous method experience. We train and optimize our model on a single 3090ti. The Adam optimizer is employed with a mini-batch size of 64 for 40 epochs. The learning rate is set to 0.0002 with a decay of 10 $\%$ every 20 epochs. The stage N is set as 3. We choose the snapshot with the best performance on the validation set for testing.

4.4 Performance Comparison

We do not use external training data to augment negative samples to enhance the contrastive learning capability and directly list the experimental results provided in the papers of each comparative method. It is worth noting that most methods provide an ensemble result, i.e., averaging similarity scores of two trained models. We also present the results of our ensemble model.

Comparisons on Flickr30K Results on the Flickr30K dataset are summarized in the left column of Table 1. Methods compared most are based on local matching through fine-grained alignment. Fine-grained alignment methods achieve better performance than coarse-grained alignment methods. This is due to their increased emphasis on cross-modal feature interactions, further bridging the semantic gap and enriching representations, albeit at the cost of lower inference efficiency. But currently due to a wide variety of sophisticated model designs and cutting-edge learning strategies, multi-grained alignment methods can achieve better results. This fact indicates the extreme

Table 2. Comparison of bi-directional retrieval results(R@K(%)) on MSCOCO 5K test set. The best performance is highlighted in bold.

Method	Text Retrieval		Image Retrieval	
	R@1	R@10	R@1	R@10
SGRAF [5]	57.8	91.6	41.9	81.3
CMCAN [26]	61.5	92.9	44.0	82.6
TERAN [13]	55.6	91.6	42.6	82.9
DIME [16]	59.3	91.9	43.1	83.1
NAAF [27]	58.9	92.0	42.5	81.4
RCAR [4]	61.3	92.6	44.3	83.2
NCAA(ensemble)	**61.7**	**93.2**	**44.7**	**84.1**

Table 3. The ablation analysis on Flickr30K to study the effect of different interaction modules and the similarity boundary μ.

Method	Text Retrieval			Image Retrieval		
	R@1	R@5	R@10	R@1	R@5	R@10
w/o CCC	77.0	92.5	95.7	58.5	84.2	89.9
w/o IMR	79.3	93.8	96.6	60.8	86.0	91.1
w/o CGE	80.1	94.6	97.2	61.5	86.6	91.9
$\mu = 0.50$	79.2	93.7	97.3	61.2	85.6	91.6
$\mu = 0.25$	81.2	94.9	98.1	62.4	87.2	**93.0**
$\mu = 0.10$	80.6	94.5	97.7	61.7	86.6	92.1
$\mu = 0.00$	80.0	93.9	97.7	61.2	86.0	91.6
NCAA($\mu = 0.20$)	**81.4**	**95.3**	**98.2**	**62.6**	**87.5**	**92.8**

essentiality of elaborately establishing interactive modules for image-text retrieval. We observe that our single model does demonstrate strong competitiveness in both text and image retrieval, surpassing the performance of many previously ensemble models in terms of the R@1 metric, the most important indicator of model performance. This verifies the effectiveness of the strategy to group semantic-related features and redundant features separately through clustering before aggregating high-order representations. Our ensemble model leads by a significant margin in all metrics except R@5. While our R@5 score is only 0.1$\%$ lower than the NAAF [27], we outperform them by a total of 8.3$\%$ in all other metrics.

Comparisons on MSCOCO The results of COCO-1K and COCO-5K are reported in the right column of Table 1 and Table 2, respectively. Our method outperforms compared baselines regarding R@K with different depths in the COCO-1K evaluation except the R@1 score of image retrieval, where it performs second-best and exclusively 0.3$\%$ lower compared with the best RCAR [4]. This once again manifests the remarkable ability of our proposed hybrid alignment stages. For the COCO-5K testing, our model continues to maintain superiority in all the evaluation metrics over previous methods, further indicating the stability and robustness of our proposed model.

4.5 Ablation Studies

In this section, we carry out several experiments on the Flicker30K using the single model to further analyze the effectiveness of our model. Specifically, we demonstrate the impact of each component of our framework on performance.

Interaction Modules To gain further insights into our three alignment modules, we conduct progressively incremental ablation experiments. We compare our single model NCAA with the following variants: 1) **w/o CCC**, excluding the CCC module; 2) **w/o IMR**, without the IMR module; 3) **w/o CGE**, removing the CGE module. As shown in Table 3, the performance drop of **w/o CCC** was the most dramatic. The $R@1$ score in both text retrieval and image retrieval demotes by 4.4% and 4.1%, respectively. **CCC**, utilizing clustering algorithms, is the most critical module among the alignment modules. This underscores the significance of considering semantic relations during the process of feature aggregation, as it enhances the discerning clues across modalities at a local level. Conversely, compared to the single model, **w/o CGE** exhibits the most modest decrease, indicating that leveraging the global information of one modality to enhance the local representation of another modality is effective. Furthermore, **w/o IMR** shows poorer performance than the single model, revealing that modality-specific reasoning and the combination of different regions from different views can strengthen the model performance. Overall, our model treating these three modules as an alignment stage is comprehensive and potent.

Stage Numbers To explore impacts of the number of feature interaction stages, we design experiments by gradually increasing the number of stages from 1 to 5. The experimental results are illustrated in Figure 6. It is evident that the retrieval performance steadily improves as the number of stages increases from 1 to 3. Enhancing the capability of the intra- and inter-modality feature interaction is crucial to address the heterogeneity between the two modalities. Increasing the number of stages enhances the frequency of feature interactions, thereby contributing to the performance improvement. However, beyond 3 stages, the increase in the number of stages has a fluctuating effect on the model, with most metrics exhibiting a downward trend, but there are still minor increases in metrics such as R@5. This phenomenon is attributed to the enlarged model complexity, which limits the optimization of the model, leading to overfitting of the feature interaction capacity and the inability to learn optimal representations. Therefore, to strike a balance between performance and efficiency, we ultimately settled on having 3 stages.

Boundary Analysis The parameter μ, serving as the boundary value, plays a critical role in our clustering method. We determined the value of μ to be 0.2 after considering the normalized similarity values in previous SOTA methods. In order to prove the validity of this boundary value and thus further prove the validity of the two masking mechanisms of matching and redundancy, we conduct a set of control experiments by setting different values, and the results are shown in Table 3. It is evident that the results are better than any set of control experiments when μ is 0.2. This strongly demonstrates the effectiveness of using two masking strategies to distinguish redundant

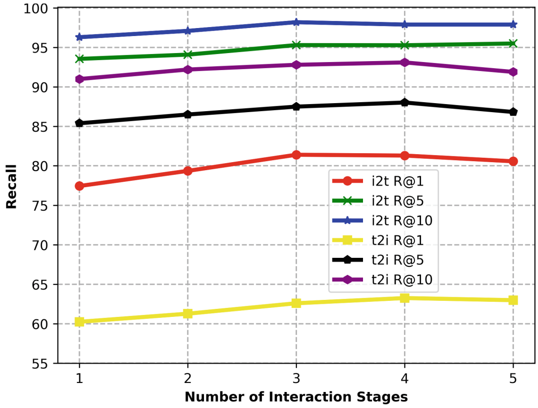


Fig. 6. results comparison on Flickr30K about number of interaction stages.

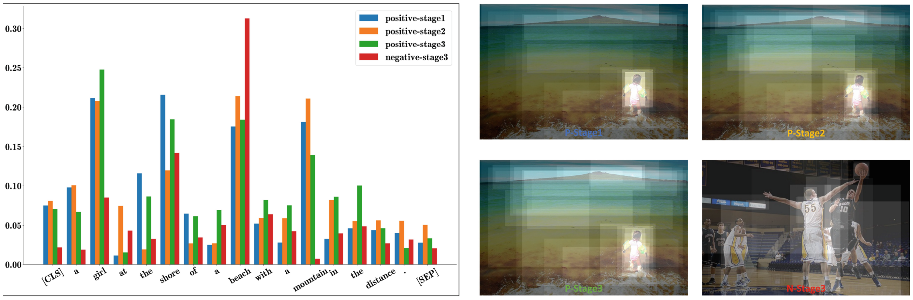


Fig. 7. The histograms display the attention weights on fragment-based alignments with the similarity attention weight while the images reflect the relative weights on region-based alignments based on the same weight.

and matching regions. Additionally, when the value of μ is larger, it means that the conditions for matching regions are more stringent, resulting in fewer regions meeting the criteria, while the conditions for redundancy are more lenient, leading to more regions being grouped as independent redundant areas, and vice versa. Therefore, performances decrease due to the unreasonable allocation of redundant or matching regions. It is worth mentioning that when the boundary value changes slightly, the performance of the model also fluctuates slightly around the optimal performance. This further presents the insensitivity and robustness of our model.

4.6 Qualitative Results and Analysis

In order to gain a more intuitive understanding of the proposed cross-modal context clustering module, we extract the attention weights w computed in this module in each stage and visualize them, as illustrated in Figure 6. It is quite evident that in instances of positive pairs, the salient regions within the image are significantly emphasized after several rounds of clustering operations. This indicates that the CCC module can selectively identify semantically matched regions while filtering out redundant ones. Similarly, the bar chart also demonstrates that alignment based on textual fragments can emphasize discriminative alignments (such as ‘girl’, ‘shore’, ‘mountain’) while discarding semantically irrelevant and redundant alignments (such as ‘a’, ‘the’, ‘at’). In the visualization of negative sample, both text histogram and region attention map reflect the tendency of model to align text and images globally rather than forcefully learning incorrect alignments. This is well demonstrated by the high values of ‘[CLS]’ in the histogram and the even distribution of attention in the regions of the image.

5 Conclusion

In this paper, we propose a Novel Cluster Aggregation and multi-grained Alignment network. Specifically, we use textual fragments as clustering centers and group regions into clusters by the local similarity. Besides, we develop two masking mechanisms to categorize regions into relevant and redundant regions in the clustering process. In order to make the model more capable of representation learning, we additionally use intra-modal multi-glance reasoning and cross-modal global enhancement. Based on these, our model NCAA can obtain intra-modal and inter-modal multi-grained features. During the inference, we also use a two-stage inference strategy. Extensive experiments demonstrate the superiority of our model. In the future, we hope to make a unified information retrieval model that makes full use of large language models(LLM) and large visual models.

Acknowledgements. This work was supported by National Natural Science Foundation of China (62020106012, 62202204), the National Key Research and Development Program of China (Grant No. 2023YFF1105102), the National Natural Science and the 111 Project of Ministry of Education of China (Grant No. B12018).

References

1. Achanta, R., Shaji, A., Smith, K., Lucchi, A., Fua, P., Süsstrunk, S.: Slic superpixels compared to state-of-the-art superpixel methods. IEEE Trans. Pattern Anal. Mach. Intell. **34**(11), 2274–2282 (2012)
2. Anderson, P., He, X., Buehler, C., Teney, D., Johnson, M., Gould, S., Zhang, L.: Bottom-up and top-down attention for image captioning and visual question answering. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 6077–6086 (2018)
3. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018)
4. Diao, H., Zhang, Y., Liu, W., Ruan, X., Lu, H.: Plug-and-play regulators for image-text matching. IEEE Transactions on Image Processing (2023)
5. Diao, H., Zhang, Y., Ma, L., Lu, H.: Similarity reasoning and filtration for image-text matching. In: Proceedings of the AAAI conference on artificial intelligence. vol. 35, pp. 1218–1226 (2021)
6. Faghri, F., Fleet, D.J., Kiros, J.R., Fidler, S.: Vse++: Improving visual-semantic embeddings with hard negatives. arXiv preprint arXiv:1707.05612 (2017)

7. Kumari, N., Zhang, B., Zhang, R., Shechtman, E., Zhu, J.Y.: Multi-concept customization of text-to-image diffusion. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1931–1941 (2023)
8. Lee, K.H., Chen, X., Hua, G., Hu, H., He, X.: Stacked cross attention for image-text matching. In: Proceedings of the European conference on computer vision (ECCV). pp. 201–216 (2018)
9. Li, J., Selvaraju, R., Gotmare, A., Joty, S., Xiong, C., Hoi, S.C.H.: Align before fuse: Vision and language representation learning with momentum distillation. Adv. Neural. Inf. Process. Syst. **34**, 9694–9705 (2021)
10. Li, K., Zhang, Y., Li, K., Li, Y., Fu, Y.: Visual semantic reasoning for image-text matching. In: Proceedings of the IEEE/CVF International conference on computer vision. pp. 4654–4662 (2019)
11. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13. pp. 740–755. Springer (2014)
12. Ma, X., Zhou, Y., Wang, H., Qin, C., Sun, B., Liu, C., Fu, Y.: Image as set of points. arXiv preprint arXiv:2303.01494 (2023)
13. Messina, N., Amato, G., Esuli, A., Falchi, F., Gennaro, C., Marchand-Maillet, S.: Fine-grained visual textual alignment for cross-modal retrieval using transformer encoders. ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM) **17**(4), 1–23 (2021)
14. Messina, N., Falchi, F., Esuli, A., Amato, G.: Transformer reasoning network for image-text matching and retrieval. In: 2020 25th International Conference on Pattern Recognition (ICPR). pp. 5222–5229. IEEE (2021)
15. Pont-Tuset, J., Uijlings, J., Changpinyo, S., Soricut, R., Ferrari, V.: Connecting vision and language with localized narratives (2019)
16. Qu, L., Liu, M., Wu, J., Gao, Z., Nie, L.: Dynamic modality interaction modeling for image-text retrieval. In: Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval. pp. 1104–1113 (2021)
17. Ren, S., He, K., Girshick, R., Sun, J.: Faster r-cnn: Towards real-time object detection with region proposal networks. Advances in neural information processing systems **28** (2015)
18. Shi, B., Ji, L., Lu, P., Niu, Z., Duan, N.: Knowledge aware semantic concept expansion for image-text matching. In: IJCAI. vol. 1, p. 2 (2019)
19. Wang, L., Li, Y., Lazebnik, S.: Learning deep structure-preserving image-text embeddings. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 5005–5013 (2016)
20. Wang, Y., Yang, H., Qian, X., Ma, L., Lu, J., Li, B., Fan, X.: Position focused attention network for image-text matching. arXiv preprint arXiv:1907.09748 (2019)
21. Yan, S., Yu, L., Xie, Y.: Discrete-continuous action space policy gradient-based attention for image-text matching. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8096–8105 (2021)
22. Yang, F., Sun, Q., Jin, H., Zhou, Z.: Superpixel segmentation with fully convolutional networks. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 13964–13973 (2020)
23. Young, P., Lai, A., Hodosh, M., Hockenmaier, J.: From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. Transactions of the Association for Computational Linguistics **2**, 67–78 (2014)
24. Zhang, D., Wu, X.J., Liu, Z., Yu, J., Kitter, J.: Fast discrete cross-modal hashing based on label relaxation and matrix factorization. In: 2020 25th International Conference on Pattern Recognition (ICPR). pp. 4845–4850. IEEE (2021)

25. Zhang, D., Wu, X.J., Yu, J.: Discrete bidirectional matrix factorization hashing for zero-shot cross-media retrieval. In: Chinese conference on pattern recognition and computer vision (PRCV). pp. 524–536. Springer (2021)
26. Zhang, H., Mao, Z., Zhang, K., Zhang, Y.: Show your faith: Cross-modal confidence-aware network for image-text matching. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 36, pp. 3262–3270 (2022)
27. Zhang, K., Mao, Z., Wang, Q., Zhang, Y.: Negative-aware attention framework for image-text matching. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 15661–15670 (2022)

Ensembling YOLO and ViT for Plant Disease Detection

Debojyoti Misra(✉), Suryansh Goel, and Tushar Sandhan

Indian Institute of Technology Kanpur, Kanpur, India
{debojyotim22,suryanshg22,sandhan}@iitk.ac.in
https://home.iitk.ac.in/~sandhan/

Abstract. As the population of the earth grows, the demand for food grows proportionally. Early and cost-effective detection of plant diseases can result in less food loss. The current methods for image-based plant disease detection tend to fail in field conditions. The proposed pipeline uses an ensemble of an YOLOv8 model trained for disease detection and a disease detection module made of YOLOv8 (as the localizer) and Vision Transformer (as the classifier). The ensembling is performed with a method called Soft-NMS. Our pipeline performs disease detection with 46.12% mAP, beating the YOLOv8 by 14.72%, which detects with 31.4% mAP in the Plant-Doc dataset.

Keywords: Plant disease detection · Object Detection · Ensemble models

1 Introduction

The world's population is currently estimated to be over 8 billion people, and it is growing by about 140 people every minute, according to the World Population Review [1]. Food security becomes a pressing concern of the hour as population grows. About 16% [2] of yield loss worldwide is attributed to plant diseases, which is a considerable amount. Increased yield will directly affect the availability of food. Farmers find it challenging and expensive to have a sample examined by a professional in order to diagnose diseases accurately. Conventional plant disease diagnosis techniques require labor-intensive visual inspection by specialists, which frequently causes responses to be delayed and may result in yield losses. Machine learning (ML) can be used to overcome this issue. The effectiveness of ML-based approaches in terms of cost and accuracy has been demonstrated with a significant amount of labelled data. Just like humans, different diseases cause distinct symptoms in plants also. One such significant indication is the patterning of leaves. This makes the identification of diseases a problem involving pattern recognition (Fig. 1).

The issue of disease identification in plants using leaf photos has been tackled by researchers in the past using a variety of strategies. One such approach [3]

A. Antonacopoulos et al. (Eds.): ICPR 2024, LNCS 15321, pp. 80–94, 2025.
https://doi.org/10.1007/978-3-031-78305-0_6

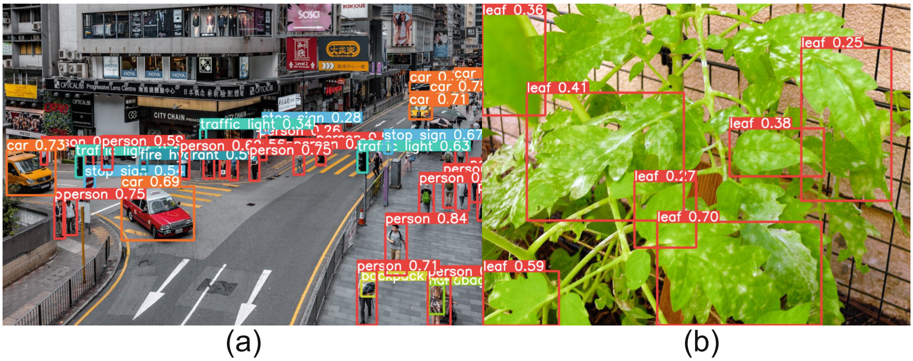


Fig. 1. In the images we can see predictions made by two different YOLOv8 [25] models in the form of rectangular boxes called bounding boxes; each box is assigned a class along with a confidence score by the model. (a)A trained YOLOv8 [25] on COCO 2017 datset for multiple objects, (b)A YOLOv8 trained on Plant-Doc [22] Object Detection dataset detect leaves.

was to directly use convolutional neural networks (CNN) which take the leaf image as input for disease classification.

The authors of a subsequent work [4] place greater emphasis on manually created features derived from the leaf image for classification. In a different line of work researchers have taken up this problem as an object detection task [5–7] where they aim to both localize and classify the diseased leaf portion. Authors have also explored transfer learning, feature fusion [8] and ensemble models [9] for this task. The ensemble model proposed in [9] is state of the art for the classification task.

The gradual improvement of YOLO's [16] performance in detection and the power of Vision Transformer (ViT) [24] in classification is the reason behind the idea of bringing them together. In this work we take the object detection approach for efficient leaf localization and propose a pipeline for disease detection consisting of object detection models and classification models. In our pipeline is an ensemble of two object detection models for disease detection. One being the YOLOv8 (multiclass) [25] and the other is the module made up of YOLOv8 (leaf) [25] and Vision Transformer.

2 Related Work

Techniques for detecting plant diseases based on vision have been around for a while. These approaches are quick and inexpensive, which makes them a great replacement for lab-based methods, which are cost-intensive and lengthy. The advancement of smartphones and the internet these days have paved the way for image-based techniques to have higher penetration in remote areas. Data being scarce in agricultural domain, it is very hard to train accurate deep learning models. Also the absence of object detection models which can perform well in smaller dataset is another obstacle for deploying object detection models in agricultural domain. For any kind of disease detection the accuracy of the method is of utmost importance. In order to be used with reliability in field circumstances, vision-based techniques must therefore be accurate.

2.1 Plant disease detection (As a classification task)

The domain of image classification has reached previously unattainable levels thanks to important developments in CNNs and other machine learning models in recent years. Particularly, CNNs have proven to be exceptionally effective at extracting hierarchical elements from images, which allows them to perform better on image identification tasks [11]. CNNs are now even more sophisticated and efficient thanks to the development of designs like ResNet, Inception, and EfficientNet [11]. Beyond CNNs, a variety of machine learning models continue to shape the field of image categorization, such as Random Forests, Support Vector Machines, and neural network variations like Capsule Networks [12].

Data being scarce in this domain, the first reliable Deep learning (DL) model we come across is from the works of Mohanty et al. [3]. They use Plant Village Dataset [10] for disease classification and attain a 99.34% accuracy rate using pretrained GoogleNet and AlexNet for transfer learning. On the other hand to make the model lite-weight and suitable for edge devices Ahmed et al. [4] use conventional image processing with ML models for classification task on Plant Village Dataset [10]. They generate handcrafted features and combine them with all sorts of ML classifiers to achieve comparable performance.In [9] the authors achieved 100% classification accuracy on Plant Village Dataset [10] by creating an ensemble of two weak models trained on two different subsets of the dataset for fast training and robust performance. Problem is that these models can't perform in field conditions. In field conditions there are multiple objects present in the image like branches,insects and soil which interfere with the disease classification. Also these models are trained on Plant Village [10] dataset which has only single leaf present in the image with clear background and good illumination.

After the introduction of ViT [24] it has given state-of-the-art performance in classification tasks on multiple datasets like ImageNet, CIFAR-10, CIFAR-100, Oxford-IIIT Pets, Oxford Flowers-102 etc. It has never been used for plant disease detection before. One important thing is that ViTs need very large amount of data to beat state-of-the-art classifiers.

2.2 Plant disease detection

As a fundamental task in computer vision, object detection has a rich history of research. Starting from the introduction of two stage region proposal based detection networks [13–15] we can see rapid development of deep learning architectures for this task. Girshick et al. first comes up with region proposal based detectors in [13] where they use selective search to propose regions and use a pretrained deep feature extractor (Convolution layers of a CNN) to generate features of the proposed regions which are then used for classifying the object present in the proposed region of interest (RoI). In his later work [14] Girshick improves speed of the network with the introduction of Fast-RCNN detector followed by his work with Ren on Faster-RCNN [15] where they use a CNN architecture for proposing regions making it a trainable part of the detection pipeline and improving both speed and accuracy. In further works we find authors proposing one stage detectors. The developement of YOLO [16] and SSD [18] show comparable results with very fast detection which can be used for realtime applications. Most recent works in object detection propose a paradigm shift where they consider it as a set prediction task first introduced by Carion et al. [19]. The detection transformer (DETR) proposed in [19] simplifies the task of object detection and is appreciated for it's end-to-end framework which is favourable for training.

Fuentes et al. in their work [7] frames plant disease detection as an object detection problem. They explore the then state of the art deep learning architectures for disease detection on their own tomato disease dataset. In the subsequent works of the same author [5], they introduce a secondary diagnosis unit comprised of several disease specific lite-weight CNNs to decrease the number of false positives. An increase in training data and number of training classes help in learning more distinguishable representations is shown by experementation in [6].

2.3 Post-processing techniques to refine bounding box predictions

Most of the the works related to object detection [5–7,13–16,18] use a post-processing technique called Non-Maximum Suppression (NMS). NMS helps in refining bounding box predictions by removing redundant bounding boxes and keeping the best one. NMS uses a strong IOU threshold to quantify redundency whereas Soft-NMS [20] uses a continuous function. Another technique proposed by Solovyev et al. [21] called weighted boxes fusion does not remove redundant boxes like NMS and Soft-NMS [20] but fuses the bounding boxes together according to their scores to create a new bounding box. These techniques can also be used to create an ensemble of different object detection models where it will remove redundant predictions from different models,keeping only the best predictions.

3 Our Method

Disease detection from leaf images is a hard task even for state of the art detectors. Though they localize well,results show that they fail frequently in classifying the diseases. To solve this, we use object detection models for leaf localization followed by a deep classifier to classify each detected leaf.

Let there are N_{cls} classes of objects which can be present in an RGB image I. The image I of resolution $H \times W$ tensor I of dimension $H \times W \times 3$. Two

YOLOv8 [25] models take the tensor I as input. One of them is trained for object detection, another one is trained for object localization.

The model trained for object detection will give an output in the form of a dictionary containing boxes, scores and labels. The values of the keys in these dictionary are tensors. Boxes are given as a tensor B of dimension $N \times 4$ made up of vectors b_i of dimension 1×4 in a format of $[x_1, y_1, x_2, y_2]$ where $i \in 1, 2, ..., N$. (x_1, y_1) is the top left vertex and (x_2, y_2) is the bottom right vertex of a bounding box b_i. N is the number of bounding boxes predicted by the model for image I. The scores are given as a tensor S of dimension $N \times 1$ where each element of the tensor $s_i \in [0, 1]$. s_i is a confidence score for the i'th bounding box. The model predicts class labels as tensor L of dimension $N \times 1$ where each element $l_i \in 0, 1, 2, ..., (N_{cls} - 1)$.

The other YOLOv8 [25] model trained for object localization is part of a detection module consisting of YOLOv8 [25] and a ViT [24] classifier. The ViTs [24] helps in classifying the objects localized by the YOLOv8 [25] model and is trained on a cropped dataset of the original dataset. From the given ground truth bounding boxes are cropped with corresponding labels to train the ViTs [24] classifier. The YOLOv8 [25] model takes the tensor I as input and generates N' predictions having boxes,scores and labels. The boxes are represented in the form of a tensor B' of dimension $N' \times 4$ made up of vectors $b'_{i'}$ of dimension 1×4 in a format of $[x'_1, y'_1, x'_2, y'_2]$ where $i' \in 1, 2, ..., N'$. (x'_1, y'_1) is the top left vertex and (x'_2, y'_2) is the bottom right vertex of a bounding box $b'_{i'}$. The scores of objectness are given as a tensor O of dimension $N \times 1$ where each element of the tensor $o_{i'} \in [0, 1]$. $o_{i'}$ is a probability for the i''th bounding box having an object. The model predicts class labels as tensor of dimension $N' \times 1$ where each element is 0. As we are trying to localize the object only, we consider the object as only one class which is represented as 0 by the model (Fig. 2).

The N' boxes, B' predicted by the YOLOv8 [25] model trained to localize are cropped out of the tensor I creating $i_{i'}$ smaller tensors warped in a predefined dimension of $w \times h \times 3$. These $i_{i'}$ tensors are sent to the ViTs [24] classifier one by one for a N_{cls} class classification where the ViTs [24] gives us probability of the object present in $i_{i'}$ in the form of a $N_{cls} \times 1$ vector. The highest probability is selected as $p_{i'}$ and a label, $l'_{i'}$ is assigned based on the position of the highest probability in the $N_{cls} \times 1$ vector. For all the N' bounding boxes $b'_{i'}$ a corresponding class label $l'_{i'} \in 0, 1, 2, ..., (N_{cls} - 1)$ and a probability $p_{i'} \in [0, 1]$ is generated. We create a tensor L' and P, both of dimension $N' \times 1$ keeping $l'_{i'}$ and $p_{i'}$as i''th components for the corresponding tensors.

The objectness score $o_{i'}$ for the i''th bounding box can be interpreted as the conditional probability of an object being present in tensor $i_{i'}$ given the tensor I. Also, the probability of the class of object present in tensor $i_{i'}$ is represented by $p_{i'}$ which can be interpreted as the conditional probability of the object belonging to class $l'_{i'}$ given the an object being present there and the input I. The score of object detection is the joint probability of an object being present in the bounding box and the object belonging to a specific class. In our case the

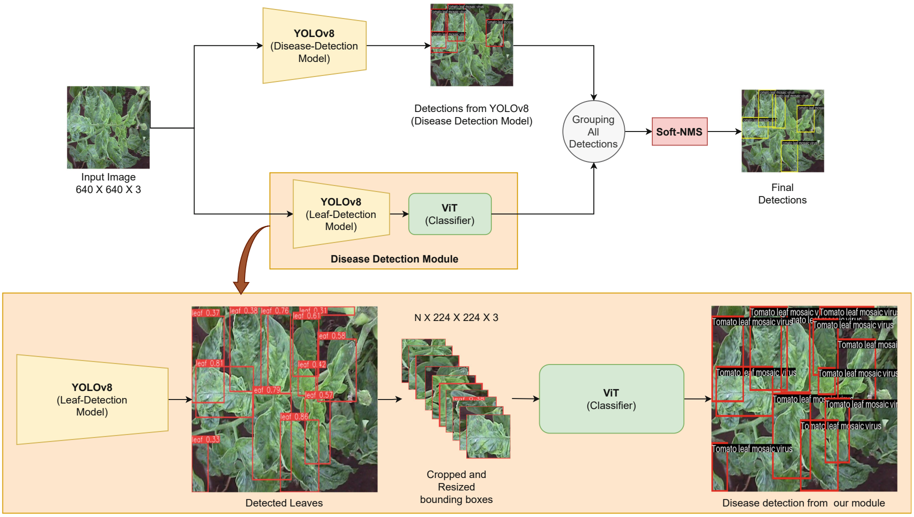


Fig. 2. Our complete pipeline for plant disease detection from leaf images. It consists of ensemble of YOLOv8 [25] and ViT [24] where we are improving the grouping of their detections via Soft-NMS [20]

score $s'_{i'}$ of the bounding box $b'_{i'}$ is $o_{i'} \times p_{i'}$ given in equation 1.

$$s'_{i'} = o_{i'} \times p_{i'} \tag{1}$$

$$S' = O \odot P \tag{2}$$

Where $\odot$ represents element-wise multiplication or Hadamard product. This provides us with a tensor S' of dimension $N' \times 1$ which represents the scores of object detection for the N' boxes present in B' tensor.

Now we have N bounding boxes from YOLOv8 [25] having boxes B, labels L, scores S and N' bounding boxes from our disease detection module having boxes B', labels L', scores S'. We group them together by concatanating them and create boxes B'' of dimension $(N+N')\times 4$, labels L'' of dimension $(N+N')\times 1$ and scores S'' of dimension $(N+N')\times 1$ for all the $N+N'$ number of bounding box predictions.

We use the Soft-NMS [20] mentioned in algorithm 1 to filter out redundant detections from the $N+N'$ detected bounding boxes. The algorithm takes two hyper-parameters α, σ and two tensors B'' and S'' as input. The bounding boxes $b''_{i''}$ are sorted according to their corresponding scores in descending order $s''_{i''}$ and the tensors B'',L'' and S'' are arranged accordingly. A IoU comparison is done between two bounding boxes to change the score of the second box accordingly. This method iteratively decreases scores for the redundant bounding boxes according to the variance of the gaussian distribution σ which decides the rate of soft score supression mentioned in the algorithm 1. The other hyper-parameter α helps us to discard the boxes $b''_{i''}$ with score $s''_{i''} < \alpha$. Soft-NMS [20]

Algorithm 1. Ensembling object detection models with Soft Non-Maximum Suppression (Soft-NMS)

```
$B, S \leftarrow$ Predictions from model 1 ; $B \in R^{N \times 4}, S \in R^{N \times 1}$
$B', S' \leftarrow$ Predictions from model 2 ; $B' \in R^{N' \times 4}, S' \in R^{N' \times 1}$
$B'' \leftarrow concatanate(B', B)$ ; $B \in R^{N+N' \times 4}$
$S'' \leftarrow concatanate(S', S)$ ; $S \in R^{N+N' \times 1}$
procedure SOFT-NMS($B'', S'', \alpha, \sigma$)
    Sort the boxes by their confidence scores $s''_{i''}$ in descending order where $i'' \in 1, 2, 3, ..., N + N'$
    $B''$ , $L''$ and $S''$ are arranged according to the descending order found in previous step
    $N + N' \leftarrow$ number of boxes
    for $i'' \leftarrow 1$ to $N + N'$ do
        for $j \leftarrow i'' + 1$ to $N + N'$ do
            $IoU \leftarrow$ Intersection-over-Union (IoU) between $B''[i'']$ and $B''[j]$
            $s''_{i''} \leftarrow s''_{i''} \cdot \exp(-\frac{IoU^2}{\sigma})$        ▷ Soft score suppression
        end for
    end for
    Apply threshold $\alpha$ to remove low-confidence boxes
    return filtered detections in form of tensors $B_P$, $L_P$ and $S_P$
end procedure
```

returns us bounding boxes B_P which have corresponding scores S_P and labels L_P which is the final output of our detection pipeline.The tensors B_P, L_P and S_P has dimensions $N_P \times 4$, $N_P \times 1$ and $N_P \times 1$ respectively where $N_P < N + N'$ (Fig. 3).

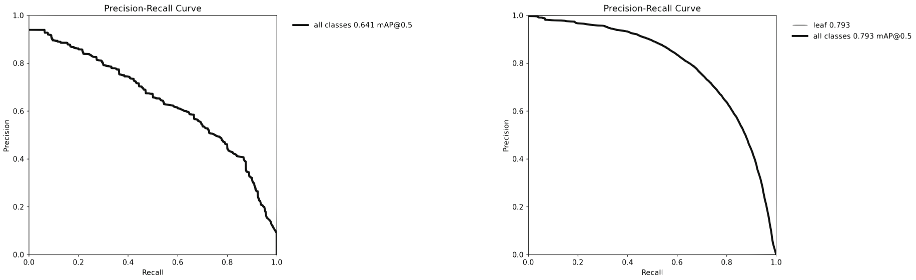
4 Experiments and Results

4.1 Experimental Setup

We have used the PyTorch framework to implement this work on a system that has an RTX 3060 GPU with 12 GB of RAM. The system runs on an Intel i7-11700K processor and 16GB of RAM.

4.2 Dataset Description

All of our studies make use of the Plant-Doc [22] dataset, which is publicly available [22]. This dataset is made up of leaf photos, the majority of which were collected in fields. Plant-Doc [22] consists of 8923 instances of leaves in a total of 2568 images. Training and testing data points are divided between 2355 and 243 for training and testing, respectively, in the dataset. There are a total of 29 classes of plant-disease pairs present in the dataset. Ground truth in the format of bounding boxes and class labels is present in the dataset, making it suitable for supervised learning.



(a)

(b)

Fig. 3. (a)Precision-Recall curve of YOLOv8 [25] disease detection Model (b)Precision-Recall curve of YOLOv8 [25] Leaf detection model showing us better and smoother convergence for localization and classification when detecting leaves in place of directly detecting diseaes

A cropped dataset is created by cropping all the labelled leaves from the original dataset. There are total 8923 cropped leaves present in the dataset. Among these we use 7351 for training, 586 for validation and 986 for testing.

4.3 Localisation

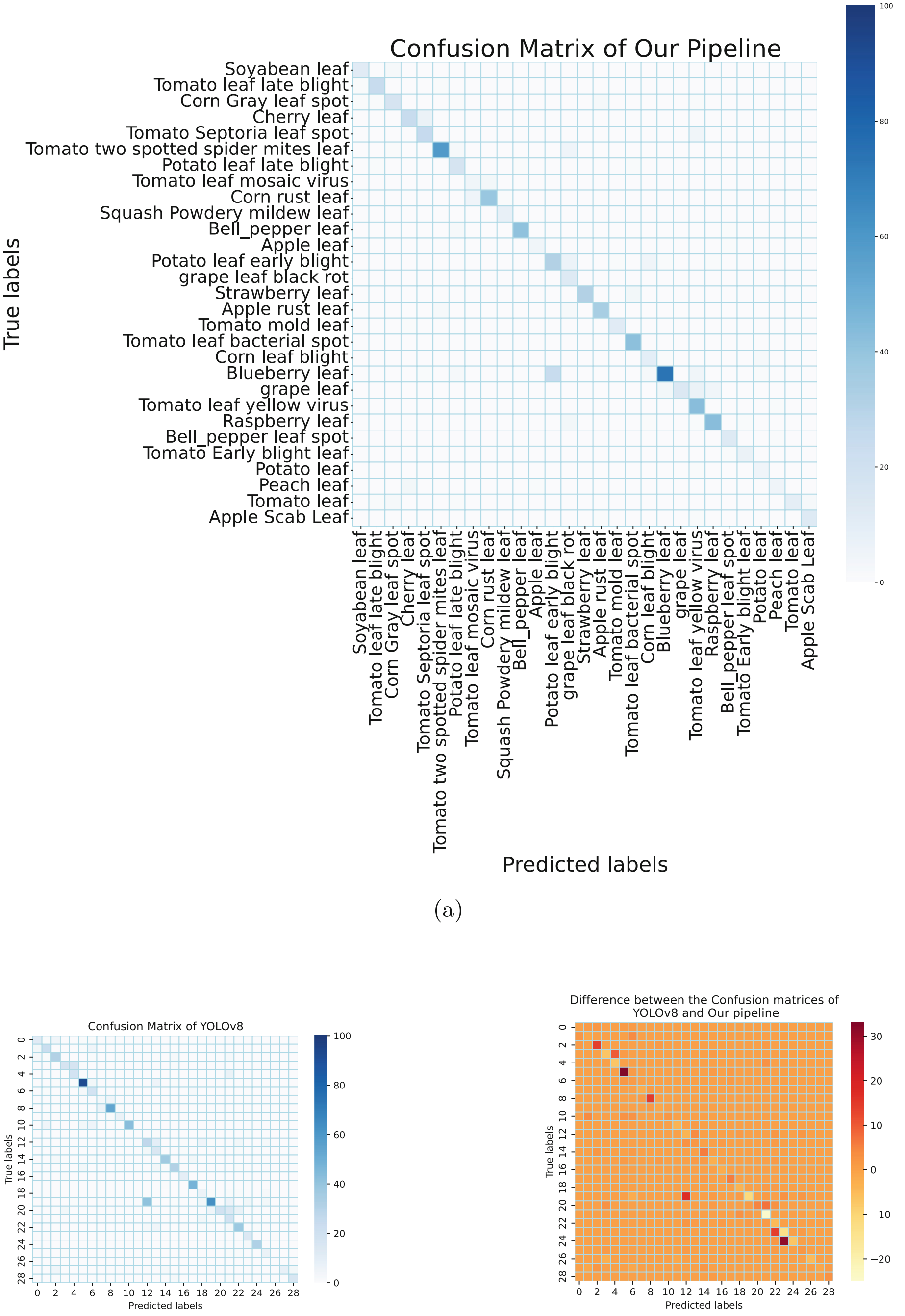
To localise the leaves efficiently, we use the YOLOv8 [25] object detection model. For this task, YOLOv8 [25] gives us a mAP of 34.09%. On the other hand, Faster-RCNN [15] reaches a mAP of 46.8%, and a fine-tuned DETR [19] fails gravely, resulting in a mAP of 8.61% as it can be seen in Table 1. Though Faster-RCNN [15] gives a higher mAP, YOLOv8 [25] is chosen due to its very fast inference time, which is in the order of milliseconds and is very suitable for creating ensembles.

Table 1. Leaf detection performancre of different state-of-the-art object detection models on Plant-Doc [22] dataset

Architecture	mAP(%)
YOLOv8 [25]	34.09
Faster-RCNN [15]	46.8
DETR [19]	8.61

4.4 Classification

In our pipeline, we assign the class of the bounding box with the help of a deep classifier. To train these models, we used the cropped Plant-Doc [22] dataset.



(a)

(b)

(c)

Fig. 4. (a)Confusion matrix of our pipeline for disease classification (b)Confusion matrix of YOLOv8 [25] for classification of diseases(c) Difference matrix between the confusion matrices of YOLOv8 [25] and our pipeline for disease classification

On the experiments performed in this dataset, we find out that Vision Transformers [24] perform the best, reaching an accuracy of 75.05% beating the VGG16 [23], VGG19 [23], Inception-V3 [26] and DenseNet161 [25] models by a large margin, where they perform inference with an accuracy of 59.47%, 57.59%, 61.93% and 69.82% respectively as reported in Table2.

In figure 4a, we can see the confusion matrix of the classification of diseases done by our pipeline. In comparison to the confusion matrix of YOLOv8 [25] seen in figure 4b, the principal diagonal in figure 4a is stronger, implying a better classification of diseases. Figure 4c is the difference between the confusion matrices of YOLOv8 [25] and our pipeline, showing us changes in detection in all the classes.

Table 2. Accuracy of different state-of-the-art classification models on cropped PlantDoc [22] dataset

Architecture	Accuracy(%)
VGG16 (trained on IMAGENET1K-V1 weights) [23]	59.46
VGG19 (trained on IMAGENET1K-V1 weights) [23]	57.59
Inception-V3(trained on IMAGENET1K-V1 weights) [26]	61.93
DenseNet161(trained on IMAGENET1K-V1 weights) [25]	69.82
ViT-b16 (trained on IMAGENET1K-SWAG-E2E-V1 weights)s [24]	73.07
ViT-base-patch16-224-in21ks [24]	75.04

4.5 Detection

When posed as a detection task where the object detection model has to localise as well as classify the object, YOLOv8 [25] reaches a mAP of 31.4%, and the Faster-RCNN [15] gives us a performance of 22.7%, which in terms of evaluation metrics is not very reliable. The plots in 3 show that the model trained for disease detection has a lower area under the curve (AUC) than the model trained for leaf detection, implying better detection performance.

For this, we propose a pipeline of YOLOv8 [25] followed by a Vision Transformers [24] classifier. This pipeline is able to reach a mAP of 26.52%, which is not at par with the detection models. But further experiments show that this pipeline is able to predict different sets of leaves present in an image other than the ones predicted by the detection models.

4.6 Ensemble of multiple object detection models

YOLOv8 [25] being very fast,it is possible to create ensembles of multiple models to increase performance without much trading off with inference time. We create an ensemble of a YOLOv8 [25] model trained for disease detection and

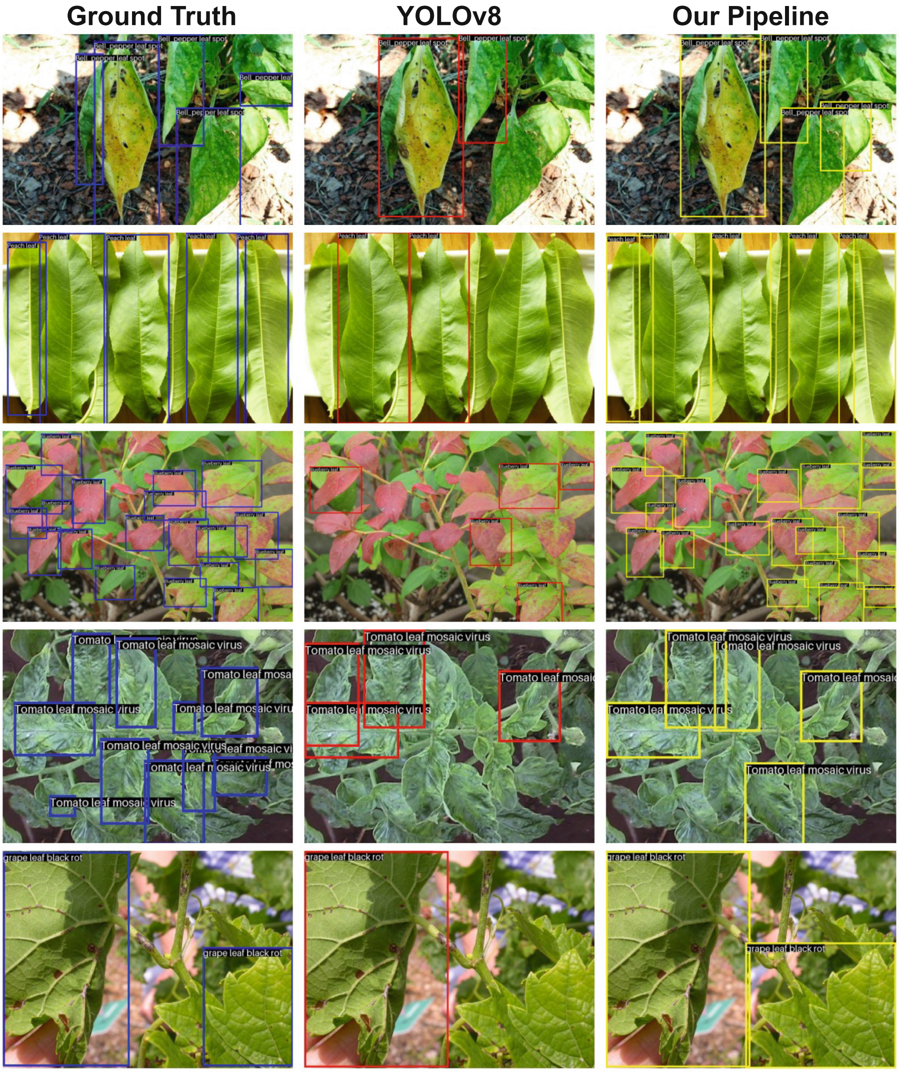


Fig. 5. In the figure above we can see three coloumns of images. In the first coloumn we can see the ground truth bounding boxes given for the corresponding image.In the second and third coloumn we can see the bounding boxes predicted by a YOLOv8 [25] model and our pipeline respectively for plant disease detection

our YOLOv8 (leaf) + ViT [24,25] pipeline trained for the same task. We take predictions from both models, put all of them together in one set, and refine them with different techniques.

These detections being refined with NMS give our ensemble model 37.4% mAP, which beats the SOTA methods by a minimum of 6% mAP. This implies that there are certain cases in the dataset that are being predicted by our YOLOv8 (leaf) + ViT pipeline but are not being predicted by the YOLOv8 [25] model; otherwise, the ensemble performance would not have improved.

On the other hand, when we apply Soft-NMS [20] in place of NMS, with threshold α=0.5 and variance of the gaussian distribution σ=0.85 ,to refine redundant detections, we can see that the ensemble model reaches a mAP of 46.12%, which beats the SOTA method by 14.72%. α and σ are hyper-parameters of the pipeline and our experimental results in table 4 show that our chosen values are optimum for ensembling. In the qualitative results shown in Figure 5, one can see that the YOLOv8 [25] misses out on localising a significant percentage of leaves present in an image during inference. In the second row of images in Figure 5, we can see that our model is able to localise leaves even from their bottom view, where the YOLOv8 [25] model is failing. In the third row, we can observe that our model gives a much higher number of predictions than the YOLOv8 [25] model, where a high number of leaves are present in the image. Even though the qualitative and quantitative results are better than the baseline YOLOv8 [25] we can see in Figure 5 that a lot of the leaves are missed when compared to the ground truth. This happens due to the Soft-NMS [20] algorithm as it suppresses low confidence detections and focuses more on accurate classification for precise object detection.

The average inference time on the test set of the Plant-Doc [22] Object Detection Dataset by the YOLOv8 [25] disease detection model, our disease detection module made up of YOLOv8 [25] and ViTs [24], and our whole pipeline are 135ms, 390ms and 401ms respectively. Which means the YOLOv8 [25] model,our module, and our proposed pipeline can process 7.4, 2.56, and 2.49 frames per second (FPS). The inference speed is a trade-off that we can accept in the task, as precise and accurate detection of the disease is more important than a very fast diagnosis (Table 3).

Table 3. Performance of plant disease detection on Plant-Doc [22] dataset

Architecture	mAP(%)
Faster-RCNN [15]	22.7
YOLOv8 [25]	31.4
YOLOv8 (leaf) + ViT-base-patch16-224-in21k (Ours) [23,25]	26.52
Ensemble (Ours) with NMS	**39.27**
Ensemble (Ours) with Soft-NMS [20]	**46.12**

Table 4. Resuls of ensembling with Soft-NMS as α and σ varies

Variance of gaussian distribution(σ)	Threshold(α)	mAP(%)
0.5	0.01	38.95
1	0.3	44.71
0.85	0.5	**46.12**
1	0.5	46.15

5 Conclusion

From the study, we can conclude that object detection models trained differently on the same dataset learn to localize differently. The YOLOv8 [25] model trained for disease detection misses out on some leaves, whereas the YOLOv8 [25] model trained for leaf detection is only capable of localising some of this. This can be clearly seen in the qualitative results.

Creating ensembles of object detection models and a secondary classifier can help increase the mAP of detection models significantly, especially in complicated tasks like ours where interclass and intraclass variation is very high. Also, it is very important to use a proper ensembling scheme to increase the performance of the model, like Soft-NMS [20].

Acknowledgements. This work is supported by IIT Kharagpur AI4ICPS I Hub Foundation, a.k.a AI4ICPS under the aegis of DST.

References

1. “2023 World Population by Country (Live).” Available:https://worldpopulationreview.com
2. Ficke, A., Cowger, C., Bergstrom, G., Brodal, G.: Understanding Yield Loss and Pathogen Biology to Improve Disease Management: Septoria Nodorum Blotch - A Case Study in Wheat. Plant Dis. **102**(4), 696–707 (2018). https://doi.org/10.1094/PDIS-09-17-1375-FE
3. S. P. Mohanty, D. P. Hughes, and M. Salathé, “Using Deep Learning for Image-Based Plant Disease Detection,” Frontiers in Plant Science, vol. 7, 2016. Available: https://www.frontiersin.org/articles/10.3389/fpls.2016.01419
4. N. Ahmad, H. M. S. Asif, G. Saleem et al., "Leaf Image-Based Plant Disease Identification Using Color and Texture Features," Wireless Pers Commun, vol. 121, pp. 1139–1168, 2021. https://doi.org/10.1007/s11277-021-09054-2
5. A. F. Fuentes, S. Yoon, J. Lee, and D. S. Park, “High-Performance Deep Neural Network-Based Tomato Plant Diseases and Pests Diagnosis System With Refinement Filter Bank,” Frontiers in Plant Science, vol. 9, 2018. Available: https://www.frontiersin.org/articles/10.3389/fpls.2018.01162
6. A. Fuentes, S. Yoon, M. H. Lee, and D. S. Park, "Improving Accuracy of Tomato Plant Disease Diagnosis Based on Deep Learning With Explicit Control of Hidden Classes," Frontiers in Plant Science, vol. 12, 2021. Available: https://www.frontiersin.org/articles/10.3389/fpls.2021.682230

7. A. Fuentes, S. Yoon, S. C. Kim, and D. S. Park, "A robust deep-learning-based detector for real-time tomato plant diseases and pests recognition," Sensors (Switzerland), vol. 17, no. 9, 2017, https://doi.org/10.3390/s17092022
8. X. Fan, P. Luo, Y. Mu, R. Zhou, T. Tjahjadi, and Y. Ren, "Leaf image based plant disease identification using transfer learning and feature fusion," Computers and Electronics in Agriculture, vol. 196, p. 106892, May 2022, https://doi.org/10.1016/j.compag.2022.106892
9. A. Bruno et al., "Improving plant disease classification by adaptive minimal ensembling," Frontiers in Artificial Intelligence, vol. 5, 2022. Available: https://www.frontiersin.org/articles/10.3389/frai.2022.868926
10. A. Pandian and G. Geetharamani, "Data for: Identification of Plant Leaf Diseases Using a 9-layer Deep Convolutional Neural Network," Mendeley Data, V1, 2019. https://doi.org/10.17632/tywbtsjrjv.1
11. L. Chen, S. Li, Q. Bai, J. Yang, S. Jiang, and Y. Miao, "Review of Image Classification Algorithms Based on Convolutional Neural Networks," Remote Sensing, vol. 13, no. 22, pp. 1–23, Jan. 2021, https://doi.org/10.3390/rs13224712
12. R. Szeliski, "Computer Vision: Algorithms and Applications, 2nd Edition"
13. R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation." arXiv, Oct. 22, 2014. Available: http://arxiv.org/abs/1311.2524
14. R. Girshick, "Fast R-CNN." arXiv, Sep. 27, 2015. Available: http://arxiv.org/abs/1504.08083
15. S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks." arXiv, Jan. 06, 2016. Availabl e: http://arxiv.org/abs/1506.01497
16. J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You Only Look Once: Unified, Real-Time Object Detection." arXiv, May 09, 2016. Available: http://arxiv.org/abs/1506.02640
17. G. Jocher, A. Chaurasia, and J. Qiu, "Ultralytics YOLO (Version 8.0.0) [Computer software]," 2023. [Online]. Available: https://github.com/ultralytics/ultralytics
18. Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C.-Y., Berg, A.C.: SSD: Single Shot MultiBox Detector. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9905, pp. 21–37. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46448-0_2
19. N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-End Object Detection with Transformers." arXiv, May 28, 2020. Available: http://arxiv.org/abs/2005.12872
20. N. Bodla, B. Singh, R. Chellappa, and L. S. Davis, "Soft-NMS — Improving Object Detection with One Line of Code," in 2017 IEEE International Conference on Computer Vision (ICCV), Venice: IEEE, Oct. 2017, pp. 5562–5570. https://doi.org/10.1109/ICCV.2017.593
21. R. Solovyev, W. Wang, and T. Gabruseva, "Weighted boxes fusion: Ensembling boxes from different object detection models," Image and Vision Computing, vol. 107, p. 104117, Mar. 2021, https://doi.org/10.1016/j.imavis.2021.104117
22. D. Singh, N. Jain, P. Jain, P. Kayal, S. Kumawat, and N. Batra, "PlantDoc: A Dataset for Visual Plant Disease Detection," in Proceedings of the 7th ACM IKDD CoDS and 25th COMAD, Hyderabad India: ACM, Jan. 2020, pp. 249–253. https://doi.org/10.1145/3371158.3371196
23. K. Simonyan and A. Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition." arXiv, Apr. 10, 2015. Available: http://arxiv.org/abs/1409.1556

24. A. Dosovitskiy et al., "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale." arXiv, Jun. 03, 2021. Accessed: Dec. 28, 2023. [Online]. Available: http://arxiv.org/abs/2010.11929
25. G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely Connected Convolutional Networks," in 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI: IEEE, Jul. 2017, pp. 2261–2269. https://doi.org/10.1109/CVPR.2017.243
26. C. Szegedy et al., "Going deeper with convolutions," in 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Jun. 2015, pp. 1–9. https://doi.org/10.1109/CVPR.2015.7298594

CIMGEN: Controlled Satellite Image Manipulation by Finetuning Pretrained Generative Models on Limited Data

Chandrakanth Gudavalli[1](✉), Erik Rosten[1], Lakshmanan Nataraj[1], Shivkumar Chandrasekaran[1,2], and B. S. Manjunath[1,2]

[1] Mayachitra, Inc., Goleta, USA
chandrakanth@ucsb.edu
[2] Electrical and Computer Engineering Department, University of California Santa Barbara, Santa Barbara, USA
https://mayachitra.com/

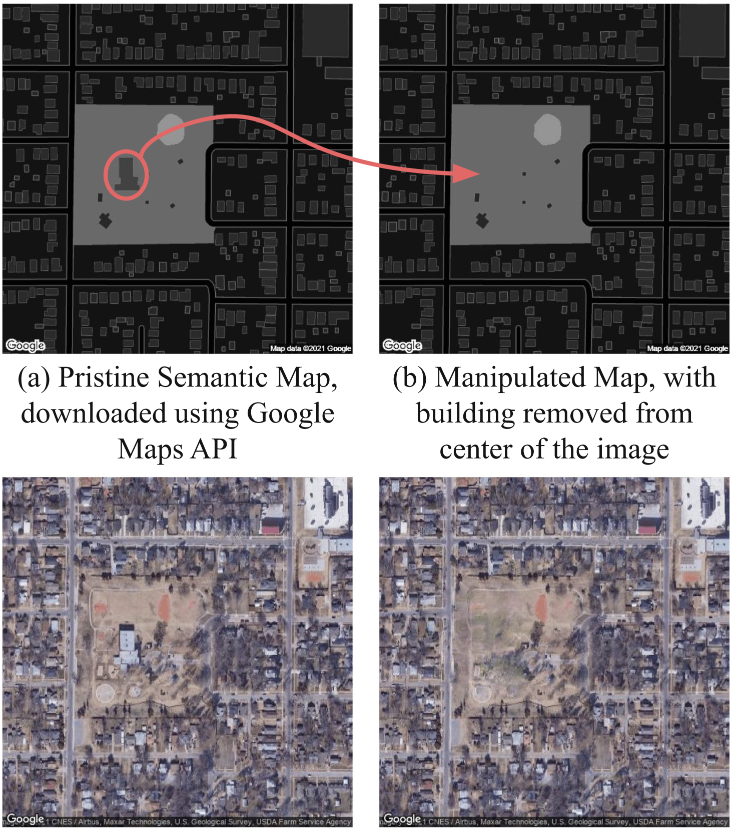
Abstract. Content creation and image editing can significantly benefit from flexible user controls. A common interpretable low-dimensional representation of an image is its semantic map, that has information about the objects present in the image. When compared to raw RGB pixels, the modification of semantic map to insert or remove objects is much easier, especially for satellite images as the satellite images are typically associated with an underlying semantic map. One can take a semantic map and easily modify it to selectively insert, remove, or replace objects in the map. The method proposed in this paper takes in the modified map of a given geographic area and alters corresponding satellite image to reflect the changes made to the map. We achieve this with traditional pre-trained image-to-image translation GANs like CycleGAN or Pix2Pix GAN, by fine-tuning them on a limited dataset of reference images associated with semantic maps. We discuss the qualitative and quantitative performance of our technique highlighting its potential for applications in satellite imagery manipulation. We also demonstrate how this method can effectively challenge numerous deep learning-based image forensic techniques, emphasizing the urgent need for robust and generalizable image forensic tools to combat the spread of manipulated data.

Keywords: Image Editing · Generative AI · Image Forensics

1 Introduction

In recent years, Generative Models (GMs) have made significant advancements in their ability to generate high-quality synthetic images and videos [4]. In the area of computer vision, these models can be applied in a variety of ways, such as generating images from text prompts [17] and performing tasks such as image to image translations. Generative Adversarial Networks (GANs) are a class of GMs that are originally introduced in 2014 [7]. One area that has seen rapid

A. Antonacopoulos et al. (Eds.): ICPR 2024, LNCS 15321, pp. 95–110, 2025.
https://doi.org/10.1007/978-3-031-78305-0_7


Google
Map data ©2021 Google
Google
Map data ©2021 Google
Google
CNES / Airbus, Maxar Technologies, U.S. Geological Survey, USDA Farm Service Agency
Google
CNES / Airbus, Maxar Technologies, U.S. Geological Survey, USDA Farm Service Agency

(a) Pristine Semantic Map, downloaded using Google Maps API

(b) Manipulated Map, with building removed from center of the image

(c) Overhead Image, that corresponds to pristine map

(d) Manipulated Image, after object removal

Fig. 1. Illustration of our proposed method, in which a building is removed from the center of an image.

growth in recent years is GAN based image-to-image translation tasks, some of which include creating art [6,11,21], inpainting images [12,20,22] and super-resolution [5,10,14]. However, generative models can also be used for more malicious intentions, producing fake images and videos. Deepfakes refer to a particular application of these types of frameworks - manipulating or synthesizing fake human faces.

In the satellite image domain, a recent paper [23] explored "deepfake geography", and the authors utilized CycleGAN [24] to transfer styles of different cities. For example, a satellite image from Seattle may be stylized to have similar landscape features as a typical satellite image from Beijing. In this paper, we present an extended, generalizable GAN based methodology that can be used for object insertion or object removal from an image (as shown in Figure 1). Our method specifically modifies (insertion/removal) specific localized regions in a semantic map, which we then *translate* to produce a fake image that agrees with the modified semantic map and resembles the original image in the unaltered regions. This generated image is then blended with the pristine image in the same localized region that the input map was manipulated, producing a doctored output image whose pixels are identical to the pristine image in the unaltered regions. These types of manipulations find potential use in applications such as defense, the agricultural sector, and urban planning. For example, in urban planning, it is quite common to have urban blueprints as maps that can undergo numerous changes in the planning stage depending on required design choices. For instance, swapping the locations of a park and a building to comply with local zoning ordinances. Using the methodology presented in this paper, these changes can be easily rendered into realistic scenes, facilitating better decision-making and public consultations. In agriculture sector for example, the map layout of an agricultural field can be modified to make way for new crops, and corresponding visualizations can be previewed in advance, aiding in efficient planning and resource management. However, the same technique can be used for malicious purposes, such as removing important structures or landmarks from aerial photographs. Therefore, it is essential to have tools for identifying such altered images. In this paper, we also discuss the limitations of a variety of deep leaning based image forensic techniques, so that the doctored images created with our proposed method cannot be flagged.

Main contributions. (1) A novel, simple, and effective way to edit or manipulate images by altering the underlying semantic map of the image. (2) Show the limits of current image forensic techniques so that fake images made with the proposed method cannot be caught.

2 Background

2.1 Generative Adversarial Networks

The GAN framework, as briefly mentioned in Section 1, consists of two neural networks that are jointly trained: a generative model G and a discriminative model D. In the most simple setup, the objective of this GAN is to generate images that are visually similar to those in the training data distribution X. However, in image-to-image translation tasks, the objective of the GAN is to learn a mapping between source domain X to the target domain Y such that $G(X)$ is indistinguishable from Y. Some example tasks include translating images from day to night, black and white to color, or map to satellite. It can be seen that D and G generally play a zero-sum game where the generator G is trying to synthesize realistic samples to fool the discriminator D. In the following subsections, we briefly cover two GAN frameworks tested with our methodology - CycleGAN [24] and Pix2pixHD [19], while noting that any architecture that can perform image-to-image translation is applicable within our proposed framework.

CycleGAN A core feature of CycleGAN [24] is its ability to learn image to image translations without paired examples. The main innovation that allows unpaired image to image translation is the addition of a *cycle consistency loss* to the objective function that is being optimized. When training a CycleGAN to learn translations between images x and y, two sets of GAN networks are trained. The first network is composed of a generator G that learns a mapping from $x \rightarrow y$, and a discriminator D_y. The second is composed of a generator F that learns a mapping $y \rightarrow x$ and a discriminator D_x.

Then, the cycle consistency loss is a measure that tries to enforce that the image translation cycle should be able to bring x back to the original image after passing it through both generators, i.e. $x \rightarrow G(x) \rightarrow F(G(x)) \approx x$. This is called forward cycle consistency. Similarly, backward cycle consistency ensures $y \rightarrow F(y) \rightarrow G(F(y)) \approx y$. We show a visual overview of this formulation for generator F in Figure 2. Mathematically, this cycle consistency term can be represented as shown in Eq 1

$$\mathcal{L}_{cycle}(G,F) = E_{x \sim p_{data}(x)}[||F(G(x)) - x||_1] + E_{y \sim p_{data}(y)}[||G(F(y)) - y||_1] \quad (1)$$

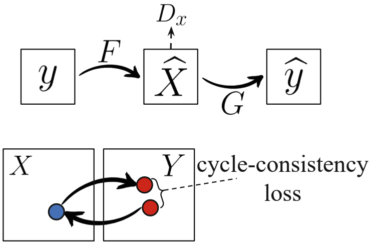


Fig. 2. CycleGAN Framework

This cycle consistency loss is added along with two unconditional adversarial loss terms, one for each generator G and F. Unconditional in this context means that the input y is not given to the generator or discriminator. This setup ensures that for a given image x, we do not need a corresponding output y to optimize the network. The full objective with a tuning parameter λ is shown in Eq 2, 3.

$$\mathcal{L}(G,F,D_x,D_y) = \mathcal{L}_{GAN}(G,D_y,x,y) + \mathcal{L}_{GAN}(F,D_x,y,x) + \lambda\, \mathcal{L}_{cycle}(G,F) \quad (2)$$

where,

$$\mathcal{L}_{GAN}(G,D,X,Y) = E_{y \sim p_{data}(y)}[logD_Y(y)] + E_{x \sim p_{data}(x)}[log(1 - D_Y(y))] \quad (3)$$

To better enforce color consistent generated images, CycleGAN also introduces an optional L_1 identity loss that encourages the generated images from each generator network match the input.

$$\mathcal{L}_{identity}(G,F) = E_{x\sim p_{data}(x)}[||F(x)-x||_1] + E_{y\sim p_{data}(y)}[||G(y)-y||_1] \tag{4}$$

In terms of network architecture, both discriminator networks take the form of a PatchGAN as in the Pix2Pix framework [9]. This PatchGAN discriminator classifies each N x N patch in an image as real or fake. Once each patch is classified, all the responses are averaged to provide a final classification from D. This patch based network network is faster to run than a full sized image classifier and is argued to take the form of a texture/style loss as it assumes that pixels separated by more than a patch diameter are independent. The generator networks are of an encoder-decoder form composed of ResNet blocks in between downsampling and upsampling layers.

Pix2pixHD The second architecture explored is Pix2pixHD , which makes several improvements over pix2pix to improve the quality of generated images that are higher resolution. The first improvement is using multi-scale discriminators and generators to ensure scene consistency at different resolution levels. The second improvement is an improved adversarial loss that incorporates a feature matching loss based on the discriminator.

In pix2pixHD, the generator G is composed of two subnetworks, G_1 and G_2 where G_1 is called a global generator network and G_2 is called a local enhancer network. Both of these networks are composed of a convolutional front-end, a set of residual blocks and transposed convolutional back-end, where the output of the global generator is half the resolution of the input image in both dimensions and the local enhancer network outputs the original input image size using the output of the global generator. During training, the global generator and local enhancers are each trained individually before being jointly trained.

The discriminator network D is also designed in a multi-scale fashion, with three discriminators D_1, D_2 and D_3 having the same network structure (PatchGAN), but D_2 and D_3 operating on 2x and 4x downsampled real and synthesized images. Then, the learning problem becomes,

$$\min_{G} \max_{D_1,D_2,D_3} \sum_{k=1,2,3} \mathcal{L}_{GAN}(G,D_k) \tag{5}$$

The second improvement in pix2pixHD is the addition of a feature matching loss. The idea here is to extract features from multiple layers in the discriminator as intermediate representations and ensure that these match for real and synthesized images. If the ith-layer feature extractor of discriminator D_k is $D_k^{(i)}$, then the feature matching loss $\mathcal{L}_{FM}(G,D_k)$ is

$$\mathcal{L}_{FM}(G,D_k) = E_{\boldsymbol{x},\boldsymbol{y}} \sum_{i=1}^{T} \frac{1}{N_i} \left[||D_k^{(i)}(\boldsymbol{x},\boldsymbol{y}) - D_k^{(i)}(\boldsymbol{x},G(\boldsymbol{x}))||_1 \right] \tag{6}$$

where T is the total number of layers and N_i is the number of elements in each layer. Then, the final full objective of the pix2pixHD GAN is

$$\min_{G}\left(\left(\max_{D_1,D_2,D_3}\sum_{k=1,2,3}\mathcal{L}_{cGAN}(G,D_k)\right)+\lambda\sum_{k=1,2,3}\mathcal{L}_{FM}(G,D_k)\right) \tag{7}$$

3 Proposed Image Manipulation Framework

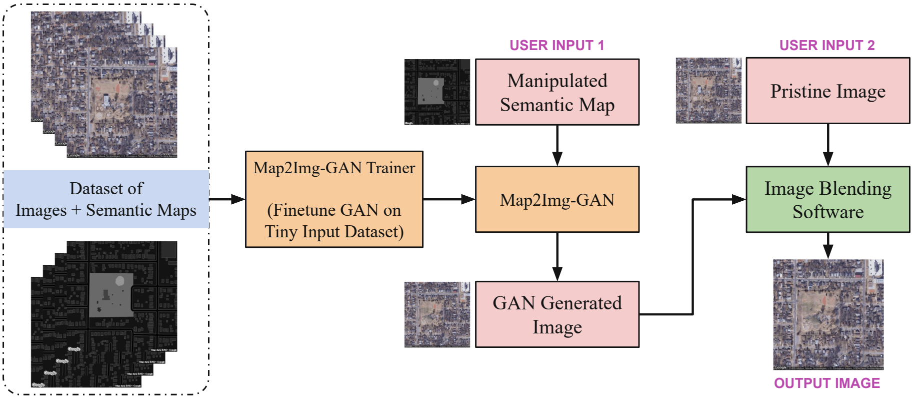


Fig. 3. Overview of proposed image manipulation framework.

We propose a simple yet effective technique for removing/replacing/inserting objects in images. We postulate using a generative model (with *traditional* CycleGAN or Pix2PiXHD GAN like architecture, pre-trained on standard large datasets in such a way that it is capable of generating images based on their semantic masks), along with a small collection of images tagged to their semantic masks. We refer to this dataset as smalldata, D, from now on. Now, we fine tune the pretrained GAN, G, on the smalldata, D, and finetune the model such that GAN generates an image that closely resembles the original image from the semantic mask that it was trained on. In other words, we mandate the GAN to memorize the correspondence between the semantic map and image. Once the GAN is trained, we generate the tampered semantic mask, T, for instance, by removing an object from the original mask, M, which is a sample in the smalldata, D. Tampering the semantic map can be easily accomplished using open-source software such as Photoshop or GIMP. Using this tampered map, T, GAN generates the image that corresponds to that tampered map, which in this case involves the generation of the image similar to the original image, that was paired with M, but with the object removed. In a similar way, one can easily insert/remove objects from the images by altering the semantic mask accordingly. This technique has produced qualitatively and quantitatively enticing outcomes, which will be discussed in detail in Section 4.

The overview of the entire framework is shown in Figure 3, where we feed a manipulated semantic map as an input to the trained generator model. For example, in the context of aerial imagery, one can remove a building and replace it with an empty land. Some applications/examples of such manipulations are discussed in detail in Section 4. Although the direct output of the generator can be used as the final result, we blend the GAN generated satellite image with the original to bolster the final manipulation's authenticity and ensure that the original pixels are preserved outside of the manipulated region. We leveraged traditional Poisson blending [15] algorithm to blend the two images, so that the pixels where semantic map is tampered will be blended into original image from GAN generated image. Possible enhancements and further research opportunities are explored in Section 6.

4 Image Forgery Experiments

4.1 Map2Sat Data Curation

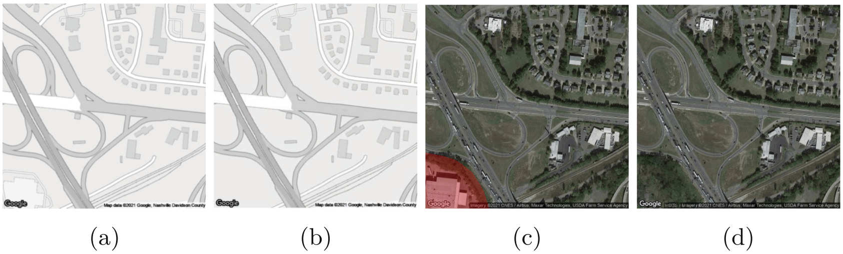
In order to curate a dataset of map to satellite imagery, we gather around 555 pairs of 512 x 512 map and satellite images of US capital cities, scraped from the google maps API. The latitude/longitude coordinates of each capital city are randomly perturbed 10 times within a 5 mile radius to obtain multiple images for each city, which differ in appearance. These resulting image pairs are manually inspected for outliers in order to constrain the domain of the dataset to urban areas. Removing these outliers brings the total number of image pairs to 470, where images were tagged as outliers for three main reasons:

- Near duplicates due to random perturbation
- Large regions of the image are not urban
- Had visual artifacts from Google stitching images together

We denote this dataset throughout the rest of the document as map2sat-urban and show a sample image pair in Figure 1a and Figure 1c.

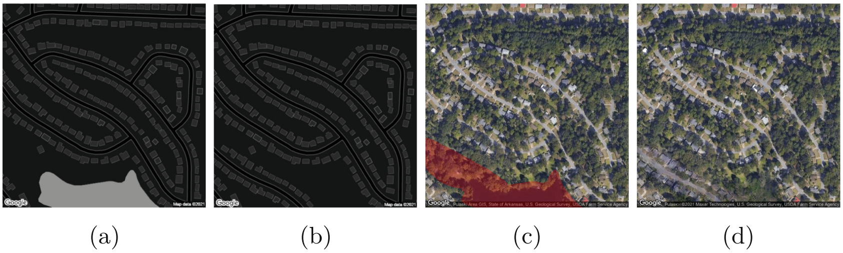
4.2 Qualitative Results

In Figure 4, we show an object removal example where we've removed a building region from the bottom left corner of the image. In Figure 5, we show an object insertion example in which a body of water is removed, and a road and buildings are inserted in its place. Both of these manipulations were generated using the method outlined in Section 3 with pix2pixHD. In Figures 6 and 7 we show manipulations generated using a trained CycleGAN model where a large cluster of buildings has been removed in the first example (Fig 6), and a larger building is removed in the third (Fig 7).


Google
Map data ©2021 Google, Nashville Davidson County
Google
Map data ©2021 Google, Nashville Davidson County
Google
Imagery ©2021 CNES / Airbus, Maxar Technologies, USDA Farm Service Agency
Google
Imagery ©2021 CNES / Airbus, Maxar Technologies, USDA Farm Service Agency

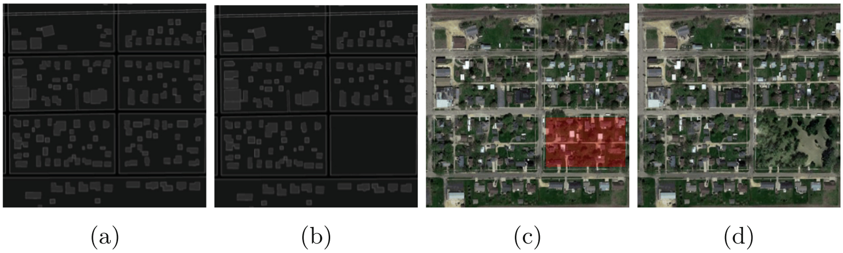
(a) (b) (c) (d)

Fig. 4. An illustration of object removal, using the proposed method with pix2pixHD GAN, from a satellite image of Nashville (the capital of the U.S. state of Tennessee). (a) Pristine Roadmap from Google Maps API. (b) Manipulated Roadmap (Objects at bottom left removed). (c) Pristine Satellite Image, overlaid on removal mask. (d) Blended Image, Objects at bottom left removed.


Google
Map data ©2021
Google
Map data ©2021
Google
Pulaski Area GIS, State of Arkansas, U.S. Geological Survey, USDA Farm Service Agency
Google
©2021 Maxar Technologies, U.S. Geological Survey, USDA Farm Service Agency

(a) (b) (c) (d)

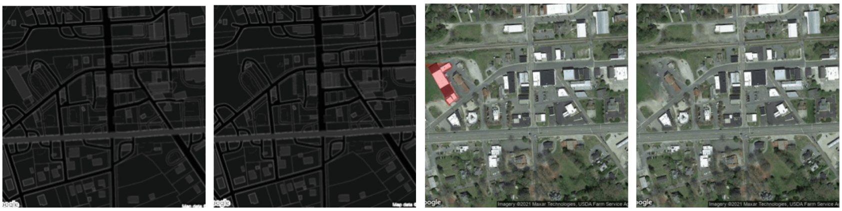
Fig. 5. An illustration of object insertion, using the proposed method with pix2pixHD GAN, from a satellite image of Little Rock (capital of the U.S. state of Arkansas). (a) Pristine Roadmap from Google Maps API. (b) Manipulated Roadmap, a water body is replaced by road and buildings. (c) Pristine Overhead Image, overlaid on insertion mask. (d) Manipulated Image (post blending)



(a) (b) (c) (d)

Fig. 6. An illustration of the proposed method with CycleGAN, where a large cluster of buildings are removed. (a) Pristine Roadmap. (b) Manipulated Roadmap. (c) Pristine Overhead Image, overlaid on removal mask. (d) Manipulated Image (post blending).

Extensions to other datasets (Building2Sat and Cityscapes): The proposed framework above can be easily extended to semantic maps of different nature than those present in map2sat-urban dataset. In addition to map2sat-urban

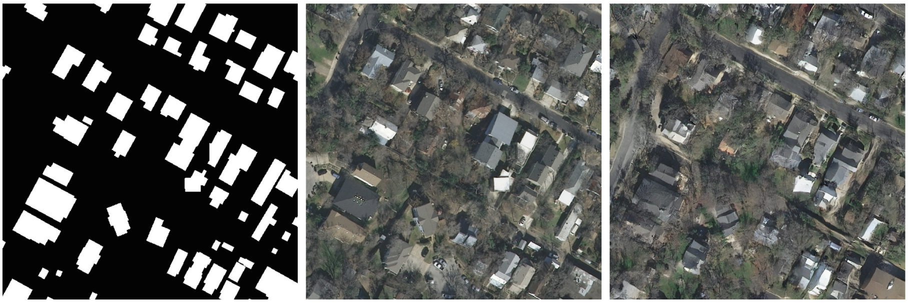


(a) (b) (c) (d)

Fig. 7. An illustration of the proposed method with CycleGAN, where a building has been removed. (a) Pristine Roadmap. (b) Manipulated Roadmap. (c) Pristine Overhead Image, overlaid on removal mask. (d) Manipulated Image (post blending).

dataset, we demonstrate our method on two other datasets: Building2Sat dataset [13] and Cityscapes dataset [3]. Building2Sat dataset is based on the Inria Aerial Image Labeling Dataset [13], which contains 180 5k x 5k image/label pairs from 5 locations: Austin, Chicago, Kitsap, Tyrol and Vienna. We select the 36 images from Austin to constrain the domain of the images, and split each 5k x 5k image into 500 x 500 tiles to obtain 3,600 image/label pairs. Utilizing pix2pixHD as our GAN, we show that our method can be used to insert/remove/add buildings in satellite images that are part of Building2Sat dataset. An example of an image/label pair along with a pix2pixHD generated image is shown in Figure 8, while Figure 9 show examples of building removal using the proposed framework.

The technique is applicable to any dataset with images and semantic maps, although it has primarily been demonstrated on aerial images thus far. The Cityscapes dataset [3], which consists of 5000 images with city street scenes labeled with extremely rich semantic maps, is one such example where we used our method. We used a subset of this dataset that has high-quality annotations, that has 5000 image map pairs to retrain the Pix2PixHD GAN. Figure 10 shows how our approach performed on this dataset with a few vehicles removed.



(a) Semantic Map (b) Pristine Image (c) Synthetic Image

Fig. 8. An example of image/label pair from Building2Sat dataset along with a pix2pixHD generated image.

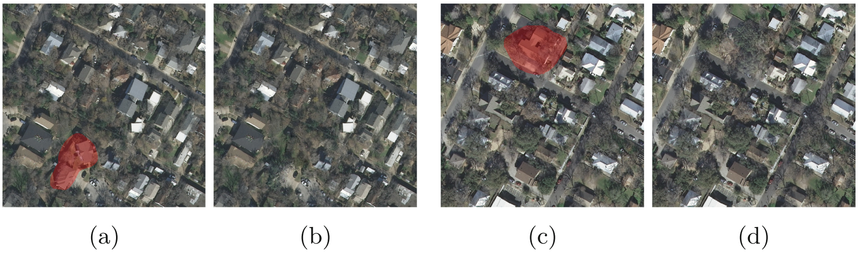


Fig. 9. An illustration of the proposed method on Building2Sat image, where a building has been removed. (a,c) Pristine Satellite Image, overlaid on removal mask. (b,d) Manipulated Image (post blending)

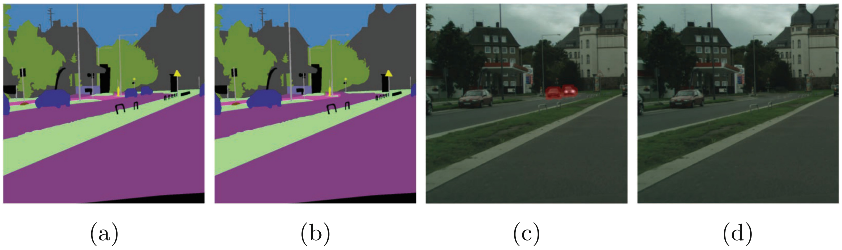
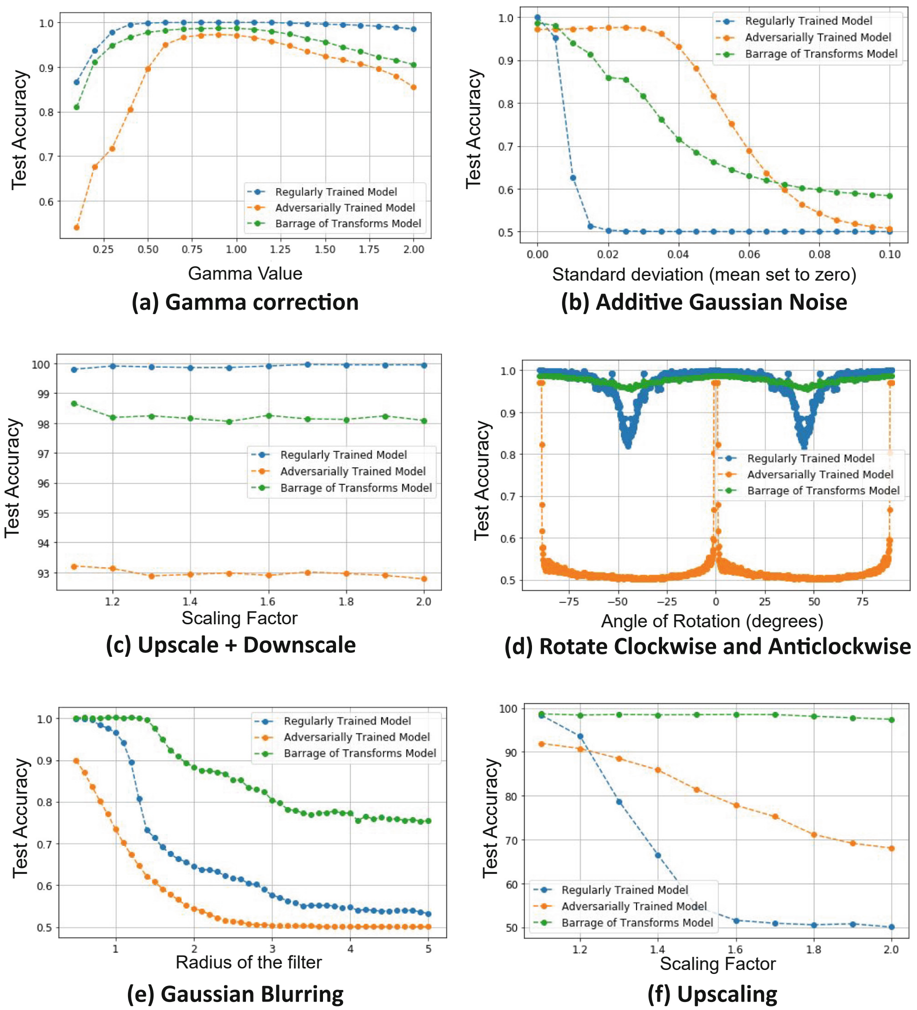


Fig. 10. Demonstration on image from cityscapes dataset, where a couple of vehicles have been removed. (a) Pristine Roadmap. (b) Manipulated Roadmap. (c) Pristine Overhead Image, overlaid on removal mask. (d) Manipulated Image (post blending).

4.3 Quantitative Results

Quantitative evaluation of the suggested technique is rather tricky due to the fact that the GAN-generated image is ultimately blended with the original image. Nevertheless, it is evident that the final blended image can only look decent if the GAN can memorize the map and image pairing and generate images with a highly similar appearance when the original map is input. To quantify this capacity of GANs, we relied on observing three standard metrics, shown in Table 1, that are typically used in literature. (1) Fréchet Inception Distance (FID) [8], (2) Kernel Inception Distance (KID) [1], and (3) Structural Similarity Index (SSIM) [18]. The following is how all three metrics are obtained. Excluding the manipulated map regions, pristine image and GAN-generated image from the manipulated map are divided into pairs of 64x64 patches. These patches are used to evaluate the FID, KID, and SSIM scores. We used patches from 20 examples in each dataset to calculate evaluation metrics. As the Build2Sat dataset has only masks for buildings, we can see that quantitative metrics are not that great on that dataset. Even though the quantitative metrics can be used for understanding the compatibility of GAN-generated images with our image manipulation framework on a very high level, manual qualitative observation is the recommended way to understand the GAN for a given application.



(a) Gamma correction

(b) Additive Gaussian Noise

(c) Upscale + Downscale

(d) Rotate Clockwise and Anticlockwise

(e) Gaussian Blurring

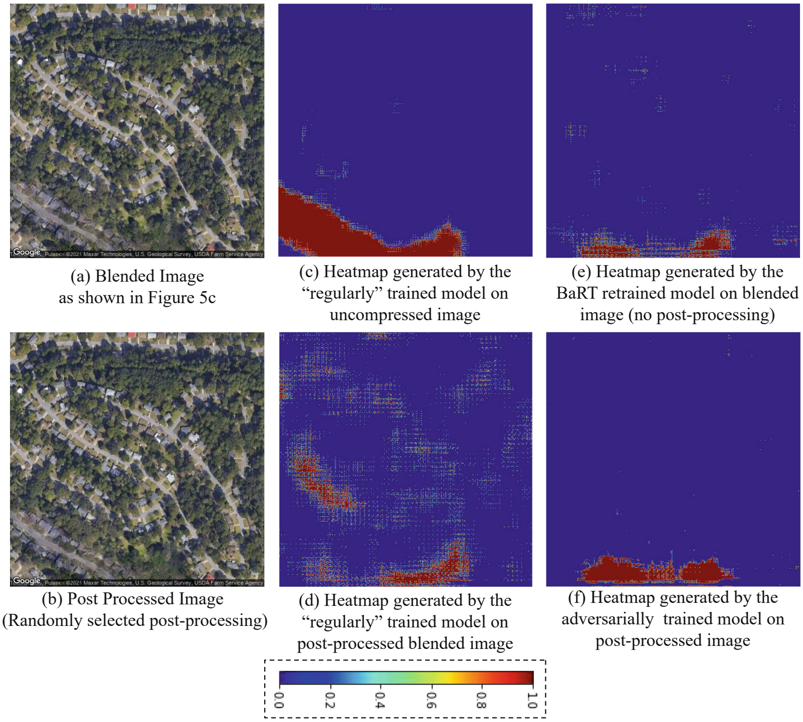
(f) Upscaling

Fig. 11. Deep learning based forensic techniques struggling to detect post-processed GAN generated images. Blue: Regularly Trained Model; Green: Barrage of Transforms Model; Orange: Adversarially Trained Model; (Color figure online)

5 Localizing Manipulated Regions

Given that trained GANs are capable of effectively forging images by removing/inserting/replacing objects, we undertook extensive experiments to determine if these manipulations may be detected by typical Convlutional Neural Network (CNN) based image forensic approaches.

The proposed framework for manipulation enables users to simply blend the GAN-generated image with the actual image. Thus, a substantial chunk of the altered image stays intact. This makes it more difficult to identify such modifi-



(a) Blended Image as shown in Figure 5c

(c) Heatmap generated by the "regularly" trained model on uncompressed image

(e) Heatmap generated by the BaRT retrained model on blended image (no post-processing)

(b) Post Processed Image (Randomly selected post-processing)

(d) Heatmap generated by the "regularly" trained model on post-processed blended image

(f) Heatmap generated by the adversarially trained model on post-processed image

Fig. 12. Deep learning-based forensic techniques have trouble finding GAN-generated images that have been changed after they were produced.

cations for image level manipulation detectors. However, it is possible to train a *patch-based classification model*, which can be a convolutional neural network-based binary classifier trained on 64x64-pixel patches. We trained a comparable patch-based ResNet50 model for detecting and localizing GAN-generated or manipulated images. Using pristine and pix2pixHD GAN-generated images

Table 1. Quantitative metrics to measure the suitability of altered images. ↓ - Lower the better. ↑ - Higher the better. The 'Pristine Baseline' represents the ideal scenario where the GAN generated patches are exactly same as pristine patches, resulting in FID and KID scores of zero, and an SSIM score of one.

Dataset	**FID** (↓)	**KID** (↓)	**SSIM** (↑)
Map2Sat-Urban	40.77	0.027	0.65
Building2Sat	81.17	0.062	0.18
CityScapes	21.44	0.006	0.73
Pristine Baseline (Ideal)	0.00	0.00	1.00

from the map2sat-urban dataset, we trained a patch-based model. Each image is divided into 64 x 64 non-overlapping patches in order to generate a dataset of 54,144 training images/patches and 6,016 validation images/patches (50 percent of them are GAN generated and the remaining half are pristine). When doing inference, we divide the input image into 64 x 64 patches and make predictions using our patch-based model with a stride of 1. With this setup, on the validation set, the trained CNN is able to achieve an AUC score of 99.99 percent and a maximum accuracy of 99.95 percent. But, we found that simple post-processing steps like rotation, scaling, gamma correction, or gaussian blurring can make the detector significantly less accurate, as shown by the blue curve in Figure 11.

Adversarial Training (AT): To train the manipulation detectors that are robust to such post-processing steps, we conducted an experiment to determine if we could improve the robustness of the detectors by adversarially training [2] the model by attacking each mini-batch during the training process with an adversarial noise under L-infinity bound of ONE. The primary purpose of the experiment is to determine whether the adversarially trained model may provide increased robustness to post-processing operations in addition to its robustness against adversarial attacks. We discovered that the adversarially trained model provides robustness to multiple post-processing stages, but at the expense of a decrease in patch-level accuracy from 99 percent to nearly 85 percent on average. Heatmaps made by models that are only 85% accurate tend to have a lot of noise, which makes them less reliable as real-time forensic detectors, as shown in Figure 12.

Barrage of Random Transform (BaRT)-based re-training [16] is another experiment that has been conducted in an effort to strengthen the forensic method. In this experiment, images were randomly post-processed on the fly while the patch-based model was being trained. The experimental setup for training the detector is explained here. With a 50 percent probability, the sample undergoes post-processing. If a sample is selected for post processing, we select one of the following post processing steps: Gamma Correction (with different gamma values), Additive Gaussian Noise (by setting mean to zero and varying the standard-deviation), Gaussian Blurring (by varying the radius of the filter from 0.1 to 5.0), Upscaling, Upscaling + Downscaling (with different scaling factors), Rotate clockwise and anti-clockwise (with varying angles of rotation).

In summary, Barrage of Random Transform (BaRT)-based re-training enables an analyst to re-train by selecting common transforms/post-processing activities. In the majority of instances, the BaRT model outperforms the regular model and the AT model, as shown in Figure 11.

On the other hand, Adversarially trained (AT) model is “universal” and “blind” to any post-processing operations with its own limitations on L-infinity bound (maximum per-pixel perturbation). It is observed that AT model outperforms regular model in some cases but is not as good as BaRT model for most cases for which BaRT model was trained on. However, strength of the adversarially re-trained model arises from the fact that the model does not know beforehand what post-processing operations that might have taken place, but it can still detect most of the operations with more accuracy relative to regular models.

AT and BaRT retraining have been proved to be more robust than naive CNN classifiers in a number of instances, however they have not yet reached accuracy levels that are regarded as reliable. However, the aforementioned forensic techniques cannot be guaranteed to be generalizable to other trained GANs, which is rather typical in our forgery pipeline given that users train GANs on smaller datasets each time they forge a new image. But, this opens the avenues for a combination of BaRT and AT models, which could prove to be far more robust.

6 Conclusion

This paper presents a framework for image manipulation with GANs. Specifically, we examined the fabrication of images from semantic maps utilizing two distinct architectures: CycleGAN and Pix2pixHD. While pix2pixHD generates images of great quality, CycleGAN has the advantage of training its generators to translate images in both directions. This advantage of CycleGAN enables manipulation of satellite images even in the absence of a semantic map, as the map can be built from the image itself using the same GAN. The methodology provided here preserves the vast majority of the original image's pixels, and can generate forgeries that are difficult for humans to identify visually. We also illustrated the ability of the proposed forging pipeline by demonstrating how it may circumvent many of the typical forensic techniques. As a future task, we are in the process of developing the similar image editing framework using other generative modeling techniques based on stable diffusion. We are also looking into efficient quantitative metrics that can employed to understand the efficacy of the pipeline. We are also exploring the possible ways to ensemble BaRT and Adversarial training strategies to make image manipulation detectors more reliable.

References

1. Binkowski, M., Sutherland, D.J., Arbel, M., Gretton, A.: Demystifying mmd gans. 6th International Conference on Learning Representations (2018)
2. Bunk, J., Chattopadhyay, S., Manjunath, B., Chandrasekaran, S.: Adversarially optimized mixup for robust classification. arXiv preprint arXiv:2103.11589 (2021)
3. Cordts, M., Omran, M., Ramos, S., Scharwächter, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S., Schiele, B.: The cityscapes dataset. In: CVPR Workshop on the Future of Datasets in Vision. vol. 2. sn (2015)
4. Dhariwal, P., Nichol, A.: Diffusion models beat gans on image synthesis. Adv. Neural. Inf. Process. Syst. **34**, 8780–8794 (2021)
5. Emad, M., Peemen, M., Corporaal, H.: Moesr: Blind super-resolution using kernel-aware mixture of experts. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 3408–3417 (2022)
6. Frühstück, A., Singh, K.K., Shechtman, E., Mitra, N.J., Wonka, P., Lu, J.: Insetgan for full-body image generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7723–7732 (2022)

7. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. Advances in neural information processing systems **27** (2014)
8. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: Gans trained by a two time-scale update rule converge to a local nash equilibrium. Advances in neural information processing systems **30** (2017)
9. Isola, P., Zhu, J.Y., Zhou, T., Efros, A.A.: Image-to-image translation with conditional adversarial networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1125–1134 (2017)
10. Ledig, C., Theis, L., Huszár, F., Caballero, J., Cunningham, A., Acosta, A., Aitken, A., Tejani, A., Totz, J., Wang, Z., et al.: Photo-realistic single image super-resolution using a generative adversarial network. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4681–4690 (2017)
11. Li, Y., Li, Y., Lu, J., Shechtman, E., Lee, Y.J., Singh, K.K.: Collaging class-specific gans for semantic image synthesis. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 14418–14427 (2021)
12. Liu, H., Wan, Z., Huang, W., Song, Y., Han, X., Liao, J.: Pd-gan: Probabilistic diverse gan for image inpainting. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 9371–9381 (June 2021)
13. Maggiori, E., Tarabalka, Y., Charpiat, G., Alliez, P.: Can semantic labeling methods generalize to any city? the inria aerial image labeling benchmark. In: IEEE International Geoscience and Remote Sensing Symposium (IGARSS). IEEE (2017)
14. Michelini, P.N., Lu, Y., Jiang, X.: edge-sr: Super-resolution for the masses. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). pp. 1078–1087 (January 2022)
15. Pérez, P., Gangnet, M., Blake, A.: Poisson image editing. In: Seminal Graphics Papers: Pushing the Boundaries, Volume 2, pp. 577–582 (2023)
16. Raff, E., Sylvester, J., Forsyth, S., McLean, M.: Barrage of random transforms for adversarially robust defense. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6528–6537 (2019)
17. Ramesh, A., Pavlov, M., Goh, G., Gray, S., Voss, C., Radford, A., Chen, M., Sutskever, I.: Zero-shot text-to-image generation. In: International Conference on Machine Learning. pp. 8821–8831. PMLR (2021)
18. Scikit-Image: Structural similarity index. https://scikit-image.org/docs/dev/auto_examples/transform/plot_ssim.html (2024), accessed: Jan 07, 2024
19. Wang, T.C., Liu, M.Y., Zhu, J.Y., Tao, A., Kautz, J., Catanzaro, B.: High-resolution image synthesis and semantic manipulation with conditional gans. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 8798–8807 (2018)
20. Wang, W., Niu, L., Zhang, J., Yang, X., Zhang, L.: Dual-path image inpainting with auxiliary gan inversion. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 11421–11430 (June 2022)
21. Xue, Y., Li, Y., Singh, K.K., Lee, Y.J.: Giraffe hd: A high-resolution 3d-aware generative model. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 18440–18449 (2022)
22. Yu, J., Lin, Z., Yang, J., Shen, X., Lu, X., Huang, T.S.: Generative image inpainting with contextual attention. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 5505–5514 (2018)
23. Zhao, B., Zhang, S., Xu, C., Sun, Y., Deng, C.: Deep fake geography? when geospatial data encounter artificial intelligence. Cartogr. Geogr. Inf. Sci. **48**(4), 338–352 (2021)

24. Zhu, J.Y., Park, T., Isola, P., Efros, A.A.: Unpaired image-to-image translation using cycle-consistent adversarial networks. In: Proceedings of the IEEE international conference on computer vision. pp. 2223–2232 (2017)

Adapt to Scarcity: Few-Shot Deepfake Detection via Low-Rank Adaptation

Silvia Cappelletti[1], Lorenzo Baraldi[2], Federico Cocchi[1,2(✉)], Marcella Cornia[1], Lorenzo Baraldi[1], and Rita Cucchiara[1]

[1] University of Modena and Reggio Emilia, Modena, Italy
`{silvia.cappelletti,federico.cocchi,marcella.cornia,`
`lorenzo.baraldi,rita.cucchiara}@unimore.it`
[2] University of Pisa, Pisa, Italy
`{lorenzo.baraldi,federico.cocchi}@phd.unipi.it`

Abstract. The boundary between AI-generated images and real photographs is becoming increasingly narrow, thanks to the realism provided by contemporary generative models. Such technological progress necessitates the evolution of existing deepfake detection algorithms to counter new threats and protect the integrity of perceived reality. Although the prevailing approach among deepfake detection methodologies relies on large collections of generated and real data, the efficacy of these methods in adapting to scenarios characterized by data scarcity remains uncertain. This obstacle arises due to the introduction of novel generation algorithms and proprietary generative models that impose restrictions on access to large-scale datasets, thereby constraining the availability of generated images. In this paper, we first analyze how the performance of current deepfake methodologies, based on the CLIP embedding space, adapt in a few-shot situation over four state-of-the-art generators. Being the CLIP embedding space not specifically tailored for the task, a fine-tuning stage is desirable, although the amount of data needed is often unavailable in a data scarcity scenario. To address this issue and limit possible overfitting, we introduce a novel approach through the Low-Rank Adaptation (LoRA) of the CLIP architecture, tailored for few-shot deepfake detection scenarios. Remarkably, the LoRA-modified CLIP, even when fine-tuned with merely 50 pairs of real and fake images, surpasses the performance of all evaluated deepfake detection models across the tested generators. Additionally, when LoRA CLIP is benchmarked against other models trained on 1,000 samples and evaluated on generative models not seen during training it exhibits superior generalization capabilities.

Keywords: Deepfake Detection · Few-Shot Learning · LoRA

1 Introduction

With the recent emergence of diffusion models [26,49] and the related enhancement in image quality, the text-to-image generative framework has facilitated the

S. Cappelletti, L. Baraldi, and F. Cocchi—Equal contribution.

A. Antonacopoulos et al. (Eds.): ICPR 2024, LNCS 15321, pp. 111–126, 2025.
https://doi.org/10.1007/978-3-031-78305-0_8

production of very realistic images from textual descriptions [5,41,42]. While this technology has enabled a wider distribution of artistic ability, it has also raised concerns about the spread of misinformation and social manipulation. To counter these side effects, deepfake detection emerges as a critical task aimed at identifying images that have been generated or altered by generative models.

Initial research in deepfake detection has mainly concentrated on identifying counterfeit faces [32,44]. Sequentially, different studies have expanded their scope to include the detection of natural images, considering a broader interest in ensuring the authenticity of a wide range of visual content. In this context, the CLIP (Contrastive Language-Image Pre-training) backbone [40] has been established as one of the most effective feature extraction methodologies for deepfake detection. Notably, when coupled with classification algorithms such as the k-Nearest Neighbor (k-NN), Support Vector Machines (SVMs), or linear classifiers, CLIP has demonstrated remarkable capabilities in discerning between generated and authentic content [1,14,36]. However, these solutions rely on large datasets comprising both real and generated images that may not be readily accessible with future generative models or commercial platforms [2,45]. Consequently, the effectiveness of CLIP-based detectors in scenarios characterized by limited data availability is still unclear and only partially approached in existing literature [14]. Further, despite the pre-trained CLIP embedding space demonstrating an ability to identify discriminative features relevant to deepfake detection, it is important to acknowledge that CLIP is optimized for a different task. For this reason, the adaptation of CLIP embedding space in the task of deepfake detection may result in improved classification results.

Low-Rank Adaptation (LoRA) [27], which originates for parameter efficient fine-tuning of large language models [16,28], has demonstrated its effectiveness in various tasks [7,8,39,48]. Specifically, LoRA allows the reshaping of an embedding space of large-scale models (*i.e.* CLIP in our scenario) by optimizing a small subset of parameters. This effect can be particularly useful in the task of deepfake detection, especially when facing scarcity in data samples, as it can effectively limit the overfitting phenomenon during fine-tuning [52]. In this paper, we conduct an experimental investigation into the few-shot learning capabilities of CLIP-based deepfake detection systems, evaluating their performance against four different state-of-the-art generative models across training sets of 20, 50, 100, and 1000 samples. Moreover, we propose a low-rank adaptation [27] of the CLIP backbone, demonstrating that efficient fine-tuning can consistently outperform other methodologies, starting with 50 pairs of real and fake images. Finally, we test the generalizable capabilities of our proposed methodology when faced with generators unseen during training, finding that LoRA reshapes CLIP embedding space toward generalized detection across different generators.

2 Related Work

Image generation models. Synthetic images are generally created using three different approaches: autoregressive models [18,21,41,57], generative adversarial networks (GANs) [6,11,29,33,50], and diffusion models [2,17,26,35,49]. Our work considers images coming from more than one family of approaches. In fact, the generated data we consider originates from Stable Diffusion [42], both the 1.4 and 2.1 versions, ProGAN [29], and DALL-E 3 [4]. To structure the image distribution, ProGAN starts with an easier task (images at low resolution) and then incrementally improves resolution step-wise while progressively adding new layers to both the generator and discriminator. Differently, Stable Diffusion models represent a specific variant of diffusion models. Indeed, these generative models operate within the latent space [34,42], augmenting efficiency while preserving the final image quality. Within the latent space, the diffusion process is conditioned through cross-attention with the U-Net layers [43]. Lastly, we consider DALL-E 3 [4], a state-of-the-art text-to-image commercial tool. This generator is available through an API and is capable of aligning images closely with the textual inputs, due to the adoption of ChatGPT [37] for prompt expansion.

Deepfake detection. The distinction between real and generated images has been an active area of research, where new classifiers are needed as generation techniques improve. Initially, detectors focused on GAN-based face generators [44,51,55]. Subsequently, with the introduction of diffusion models, detectors rapidly adapted to natural images, expanding the horizons of the face domain [1,3,13,20]. Differently from analyzing RGB data, some approaches [13,22] have utilized frequency analysis, as the generated images show spectral features that differ from real ones. Moreover, a different approach to diffusion models is explored by Wang *et al.* [54], who works on the difference between the input image and the one reconstructed by a pre-trained diffusion model.

Within the domain of deepfake detection, a significant challenge is the adaptation to generators not encountered during training, which tests the ability of the model to generalize. Recent approaches respond to this issue by employing CLIP as a pre-trained backbone from which to extract visual features used for deepfake detection [1,14,36,47]. Notably, these approaches do not use the semantic properties derived from the alignment of text and image during pre-training; rather, they leverage distinctive patterns extracted from the visual backbone. Subsequently, these visual features are utilized by classifiers to execute a binary classification task. Classifiers that have been explored in this context include Support Vector Machines (SVMs) [14], linear classifiers [1,36], and k-NN [36]. While the visual features extracted from the pre-trained CLIP embedding space are not specifically trained for deepfake detection, our approach employs LoRA fine-tuning [27] for remodeling the embedding space of CLIP with a small number of samples, with the final goal of improving deepfake classification.

3 Proposed Method

In this paper, we focus on the task of distinguishing real images (*i.e.* captured via photographic devices) from those completely generated through AI systems. In the existing literature, methodologies tackle this challenge through the creation of extensive datasets, considering thousands of real and fake images. In contrast, our research explores a distinct scenario where the availability of images from each generator for the training phase is significantly constrained. This assumption is validated in a real-world context wherein a newly released generator is unlikely to publish extensive samples, thereby restricting the availability of data for training purposes. Similarly, for closed-source generators large quantities of images are not publicly available.

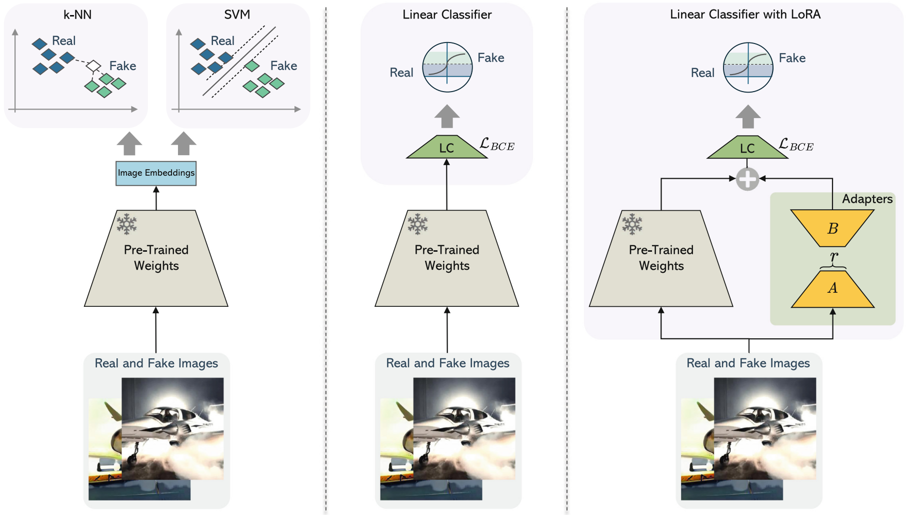


Fig. 1. Illustration of the evaluated deepfake detection classifiers. On the left, images are processed through a pre-trained backbone, with k-NN and SVM classifiers being fitted on the resulting image embeddings. In the center, a linear classifier (LC) is added on top of the backbone and trained using binary cross-entropy. On the right, our proposed fine-tuning protocol using LoRA; where LoRA adapters are added with pre-trained weights and concurrently trained with a linear classifier.

3.1 Preliminaries

CLIP architecture [40] has been recently applied in the realm of deepfake detection. Indeed, the visual features extracted from this large-scale model have been proven discriminative in this task, leading to the introduction of multiple binary classifiers (*i.e.* k-NN, SVMs, and linear layers) added on top of the frozen CLIP backbone to perform the task of fake detection [1,14,36].

Employing CLIP for a few-shot classification task offers the advantage of preventing the necessity for initial training. However, the CLIP model was originally developed with a distinct objective, *i.e.* the optimization of image-text similarity. Consequently, the ability of the CLIP embedding space to differentiate between synthetic and authentic images emerges as a secondary function of the architecture, prompting us to adapt the embedding space specifically for the task of deepfake detection.

Given the constraints of few-shot scenarios and building on the hypothesis that features relevant to deepfake detection occupy a compact subspace within the CLIP embedding domain, we investigate the efficacy of LoRA [27] in addressing this issue. In Fig. 1, we represent detectors that leverage image embeddings extracted from a pre-trained backbone. In particular, on the right, the adoption of low-rank adaptation is illustrated.

3.2 LoRA for Deepfake Detection

Given a collection of real images R and fake images F, generated from a specific deepfake generator, we select a small collection $\{(F_i, R_i), i \in (1, N)\}$ of N pairs each composed of a fake image F_i and a real image R_i. Images are firstly cropped to a size of 224^2 and then normalized by a pre-processing pipeline. Secondly, a CLIP visual backbone is employed for feature extraction. Instead of maintaining the weights frozen, we introduce trainable matrices (*i.e.* LoRA adapters), based on rank decomposition and applied into every linear layer of the backbone.

From a mathematical perspective, given a rank r and an initial weight matrix $W_0 \in \mathcal{R}^{d \times k}$, where $r << \min(d, k)$, LoRA introduces a novel formulation for weight matrices as delineated by:

$$W = W_0 + \frac{\alpha}{r} BA. \tag{1}$$

Here, $B \in \mathcal{R}^{d \times r}$ and $A \in \mathcal{R}^{r \times k}$ represent the matrices introduced for adaptation. Throughout the training process, the original weight matrix W_0 remains frozen, while B and A are optimized. Following the original implementation, B is initialized with zeros, and A is initialized from a Gaussian distribution. Conversely, α functions as a hyperparameter, modulating the degree of influence imposed by the matrices introduced by LoRA.

After CLIP processing, each R_i and F_i image is embedded in CLIP embedding space and represented as $\mathcal{F}_{R_i}$ and $\mathcal{F}_{F_i}$ features. A binary linear classifier (LC) is then trained to separate these features into distinct classes through a binary cross-entropy loss. The employment of LoRA adapters facilitates the reshaping of the CLIP embedding space, to separate $\mathcal{F}_R$ and $\mathcal{F}_F$ in the low-rank subspace. A better feature separation would result in an improved classification boundary between real and fake data.

Notably, training images R_i and F_i are chosen to represent the same semantical content. This is done to avoid a real-fake separation inside the CLIP embedding, based on the semantical properties of extracted features.

In Table 1, we detail the number of trainable parameters for each examined LoRA configuration. Significantly, the most extensive configuration encompasses 25M parameters, which corresponds to merely 7% of the Vision Transformer Large model (ViT-L) [19] employed in our experiments. During the evaluation phase, the trainable parameters are combined with the frozen weights of the backbone, as seen in Equation 1. This procedure does not result in an increase in computational load during the inference phase, as the number of parameters remains the same as in the original model.

Table 1. Number of trainable parameters for each examined LoRA configuration and linear classifier (LC) baseline. At evaluation time, adapters and pre-trained weights are merged resulting in the same number of parameters of CLIP LC.

Model	r	α	Trainable Parameters
CLIP LC	-	-	3k
LoRA CLIP LC	16	32	6M
LoRA CLIP LC	32	64	12M
LoRA CLIP LC	64	128	25M

Notably, a reduced rank r reflects in the update of a smaller number of parameters, making it advantageous for training processes that involve limited data. However, this scenario imposes a dimensional limit on the deepfake subspace within the CLIP embedding space; an increased rank may alleviate this limitation. Conversely, fine-tuning the whole visual backbone could face two drawbacks. First, fine-tuning all parameters on a small quantity of data could highly induce the overfitting phenomenon. Second, by completely redefining the CLIP embedding space, it is possible to lose the generalization capability of the network to unseen generators during training.

4 Experiments

In this section, we first describe the evaluation protocol detailing the training data, the backbone used, the baselines employed for the experiments, and the implementation details. Subsequently, we conduct experimental investigations on our proposed LoRA methodology and competitors across various state-of-the-art generative models. Within this context, we consider variations in the number of samples and analyze the generalization capabilities on unseen generators.

4.1 Evaluation Protocol and Experimental Setting

Datasets. The study of the few-shot detection capabilities of deepfake detectors requires an analysis across various deepfake generative methods. This necessity stems from the assumption that different generators may exhibit divergent behaviors in a limited-sample context, thereby requiring a varying quantity of samples to achieve acceptable detection performance. Through our experimentation, we analyze four different state-of-the-art generators, namely ProGAN, Stable Diffusion v1.4, Stable Diffusion v2.1, and DALL-E 3.

In particular, ProGAN [29] represents a popular GAN generator trained on the LSUN dataset [56], which has been deeply analyzed in the context of deepfake detection [36,53]. Differently, Stable Diffusion v1.4 (SD 1.4) and Stable Diffusion v2.1 (SD 2.1) [42] consist of two open-source diffusion models trained on the LAION dataset [46] for text-to-image conditioned generation. Finally, DALL-E

3 [4] represents the latest commercial tool introduced by OpenAI in the field of diffusion models applied to image generation.

We consider a total of 728k images from the collection introduced by Wang *et al.* [53], which includes fake images generated with ProGAN and real images coming from the same LSUN [56] classes as the generated ones. We generate nearly 14k images for both SD 1.4 and SD 2.1. This generation is performed by collecting 14k real images associated with a textual prompt from the LAION-400M dataset [46], which are then used as conditioning text to the diffusion models. Regarding the DALL-E 3 generator, we obtain 10k images from a publicly accessible collection[1]. Given the absence of corresponding real images in the dataset, we combine DALL-E 3 images with randomly selected real images from LAION-400M dataset. From these data sources, we consider 4k and 1k real-fake image pairs, respectively to create the test and validation sets for each of the four considered generators. Moreover, concerning the training set, we sample N pairs of images from the data collection to explore various few-shot scenarios, as will be introduced in Section 4.2.

Given the significant influence of image compression in the context of image forensics [12,24] and acknowledging that the majority of real images from the LAION dataset are encoded in JPEG format, we standardize all images by converting them to JPEG. This ensures uniformity in the dataset, thereby mitigating any potential bias related to varying image compression formats.

Backbone and deepfake detectors. As previously introduced, we primarily focus on the CLIP backbone. Specifically, we employ the CLIP ViT-L model pre-trained on the DataComp dataset [23] and explore different classifiers added on top of the network, such as k-NN, SVM, and linear classifiers.

Following previous literature [36], we implement a k-NN classifier, setting $k = 3$ and employing cosine distance. In this case, a feature bank is constructed by processing the training images and storing the extracted features. During the evaluation phase, the class of an image is determined by identifying the three feature vectors within the bank that exhibit the highest cosine similarity to the feature vector of the given example image. Distinctly, another baseline introduces a Support Vector Machine (SVM) classifier with a linear kernel, adopting the approach proposed by Cozzolino *et al.* [14]. Both k-NN and SVM classification processes are depicted on the left side of Fig. 1. Furthermore, we construct an additional classifier by integrating a linear classifier (LC) for binary classification on top of the CLIP backbone. This deepfake classifier is trained with binary cross-entropy loss, and a threshold of 0.5 is employed for separating real and fake images. Following previous research efforts [20,53], we additionally conduct experiments using a ResNet50 architecture [25] pre-trained on ImageNet and combined with a linear classifier.

Differently, our proposal consists of adding LoRA adapters to all the ViT-L linear layers (*i.e.* multilayer perceptron and attention layers). In our experiments, we apply the adapters only on the weight matrices, excluding the biases, and maintain a constant ratio of α to r, fixed at a value of 2 to balance adaptation

[1] https://huggingface.co/datasets/OpenDatasets/dalle-3-dataset

Table 2. Accuracy results when training with 20, 50, and 100 pairs of real and fake images and testing on the same generator. The results represent the average on five different runs with different pairs of images.

	ProGAN			**SD 1.4**			**SD 2.1**			**DALL-E 3**		
Model	20	50	100	20	50	100	20	50	100	20	50	100
ResNet50 LC	50.2	50.5	50.5	51.0	51.6	51.7	50.3	51.1	51.2	52.0	52.0	52.2
CLIP k-NN	62.7	65.6	68.0	56.8	57.0	57.7	57.2	58.2	59.2	59.3	63.3	68.6
CLIP SVM	**88.5**	91.6	93.2	**69.8**	73.8	76.1	**70.8**	75.5	76.6	**87.7**	**90.1**	91.5
CLIP LC	85.8	90.8	92.6	68.3	73.7	76.7	68.8	74.9	77.7	82.9	88.4	91.0
LoRA CLIP LC	88.1	**93.2**	**96.0**	69.4	**75.4**	**79.7**	70.2	**76.8**	**79.5**	82.5	89.7	**92.1**

and stability. Additionally, our configuration leverages a linear classifier on top of the backbone.

Implementation details. With a limited number of training samples, the use of image transformations emerges as a critical operation to mitigate the risk of overfitting. As a consequence, we select various types of image transformations, including blur, brightness, aspect ratio, pixelization, rotation, contrast, saturation, encoding quality, opacity, overlay stripes, pad, scale, sharpen, skew, grayscale, and horizontal flip. During training, each image is subjected to a stochastic process where the number of transformations applied is randomly selected from a range between 0 and 2. This approach is designed to introduce controlled variability into the training data without visually compromising images by applying too much data transformation. Moreover, to increase the variability of our data, each chosen image transformation is applied with a random strength value. This is sampled from five equally spaced ranges, generated by dividing the interval between a minimum and maximum value that we set for each transformation, with the aim of maintaining visual consistency and usability. With this configuration, we obtain five unique variants for every transformation, each with a different bounded level of intensity. Considering this random selection, all training images undergo random cropping to a dimension of 224^2. Conversely, during the evaluation phase, only a center-crop transformation is applied, at 224^2. Following this pre-processing step, each image is processed by a visual backbone. Specifically, when employing the CLIP, feature extraction is conducted from the next-to-last layer, following [14]. This approach avoids the final linear projection into the shared image-text CLIP embedding space.

From a technical standpoint, model training is performed with batch size 16, a learning rate set to $1e^{-3}$, and the SGD optimizer. The training consists of a maximum of 150 epochs, while the learning rate is reduced by a factor of 10 whenever no validation accuracy improvement is faced in the last 10 epochs. Training is automatically stopped if the learning rate reaches $1e^{-7}$. Considering the limited volume of training samples typically encountered, the evaluation phase is scheduled to occur after every two epochs of training, thereby optimizing computational efficiency.

4.2 Experimental Results

We evaluate the performance of deepfake detectors across a variety of few-shot scenarios. In particular, detectors are trained on varying numbers of pairs of samples (real and fake) N, specifically 20, 50, 100, and 1000. Considering the limited sample size in scenarios where $N \in \{20, 50, 100\}$, we conduct the experiments across five distinct random seeds, reporting the average results. This operation allows the selection of diverse sets of image pairs for each iteration, to maximize the robustness of the experimental configuration. Conversely, in the case where $N = 1000$, the results from a single random seed are reported, given that the increased number of samples inherently guarantees better stability.

Table 3. Accuracy results when training with 1000 pairs of real and fake images and testing on the same generator.

	ProGAN	**SD 1.4**	**SD 2.1**	**DALL-E 3**
Model	1000	1000	1000	1000
ResNet50 LC	50.4	52.0	51.4	52.6
CLIP k-NN	76.3	62.7	65.9	79.3
CLIP SVM	96.6	82.2	79.7	91.8
CLIP LC	97.4	83.4	84.3	92.8
LoRA CLIP LC	**99.5**	**94.1**	**90.7**	**95.7**

Evaluation on few examples. In Table 2, we report the accuracy results of our LoRA-modified CLIP model in comparison to other deepfake classifiers, specifically in scenarios characterized by a limited number of examples, namely $N \in \{20, 50, 100\}$. Notably, the efficacy of the detection models varies across different generative models. For example, employing CLIP with an SVM classifier yields accuracies of 88.5% and 87.7% for ProGAN and DALL-E 3, respectively, with $N = 20$. However, the accuracy diminishes to 69.8% and 70.8% when applied to SD 1.4 and SD 2.1, respectively. Similarly, with $N = 50$, our LoRA-enhanced model achieves accuracies of 75.4% and 76.8% for SD 1.4 and SD 2.1, respectively, whereas the results are notably higher for ProGAN and DALL-E 3, standing at 93.2% and 89.7%. This variance in performance is attributed to the various representations of different generators within the CLIP embedding space, resulting in the importance of evaluating few-shot accuracy across a spectrum of different types of generators.

When analyzing the effectiveness of detection strategies, it is noticeable that the LC paired with ResNet50 underperforms. For instance, this classifier achieves a mere 2.2% improvement in accuracy compared to random choice accuracy, *i.e.* 50%, on DALL-E 3 with $N = 100$. Differently, in the same configuration, CLIP combined with LC obtains an accuracy of 91%. This proves the effectiveness of leveraging large-scale models for few-shot deepfake detection.

Comparing our LoRA detector with the classifiers, it is evident that while performance is comparable with $N = 20$, our proposal obtains the best results with $N = 50$ and $N = 100$. For instance, LoRA CLIP obtains 93.2% and 79.7% with $N = 50$ and $N = 100$ respectively on ProGAN and SD 1.4, obtaining a gain of 1.6% and 3.6% over the SVM mode. Also, our solution demonstrates superior performance compared to the baseline CLIP LC in the majority of comparisons. Notably, even with a smaller sample size $N = 20$, our model surpasses this competitor across ProGAN, SD 1.4, and SD 2.1, with accuracy improvements of 2.3%, 1.1%, and 1.4% respectively. This indicates the efficacy of adapting the CLIP embedding space for deepfake detection even with minimal data availability, underscoring the adaptability of our proposal in a few-shot scenario.

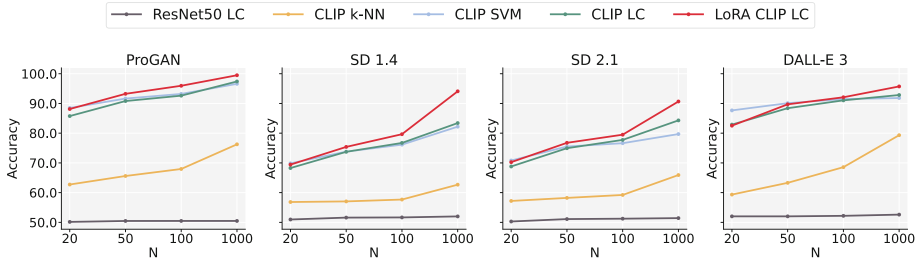


Fig. 2. Trend of accuracy scores on multiple generators. Each classifier is trained on different numbers of samples N, with $N \in \{20, 50, 100, 1000\}$ and tested on the same generator. An accuracy of 0.5 indicates that performance is equivalent to random choice.

Evaluation on more examples. In Table 3, we detail the accuracy scores of detectors, now evaluated in the context of $N = 1000$ sample pairs. Although we still consider this a few-shot scenario, it presents a relaxed constraint compared to the previous analysis.

Our LoRA-enhanced CLIP model surpasses all competitor models across all generators. Remarkably, our approach achieves accuracy improvements of 2.1%, 10.7%, 6.4%, and 2.9% over the CLIP LC model, which is the second most effective in this comparison. Further, CLIP equipped with a linear classifier demonstrates superior scalability in the $N = 1000$ scenario compared to the SVM classifier across all generators, showing performance gains of 0.8%, 1.2%, 4.6%, and 1% for ProGAN, SD 1.4, SD 2.1, and DALL-E 3, respectively.

Finally, while the performance of all methods across all generators tends to increase, as expected, with the increase of N, our proposed model demonstrates a more pronounced improvement in response to the increment of N. This trend is visually delineated in Fig. 2, providing evidence of the scalability of our approach when training size increases.

Effects of hyperparameters r and α on LoRA performance. In Table 4, we report an ablation study on the LoRA hyperparameter rank r when tested on ProGAN and SD 1.4 generators. Notably, the accuracy scores of the LoRA

CLIP model show a positive correlation with the hyperparameter r. Specifically, within the context of SD 1.4 and a sample size of $N = 50$, $r = 16$ obtains an accuracy of 73.4%, while using $r = 32$ and $r = 64$ reach an accuracy of 74.9% and 75.4% respectively. Moreover, considering $N = 20$, the $r = 64$ configuration performs better than $r = 16$ on both ProGAN and SD 1.4 with accuracy gains of 2.6% and 0.5%. This performance improvement is particularly remarkable given the substantial increase in learnable parameters, nearly 20M, associated with the $r = 64$ configuration compared to $r = 16$. Moreover, across all configurations of r, LoRA models demonstrate superior performance in comparison to the baseline CLIP LC model, proving the validity of our introduced approach independently by the analyzed hyperparameter choice.

Table 4. Accuracy results of different LoRA configurations when training with 20, 50, 100, and 1000 pairs of real and fake images and testing on the same generator. When considering 20,50, and 100 samples, the results represent the average on five different runs with different pairs of images.

			SD 1.4				**ProGAN**			
Model	r	α	20	50	100	1000	20	50	100	1000
CLIP LC	-	-	68.3	73.7	76.7	83.4	85.8	90.8	92.6	97.4
FT CLIP LC	-	-	65.3	69.9	77.2	**96.1**	52.6	54.8	60.5	99.1
LoRA CLIP LC	16	32	68.9	73.4	77.5	88.9	85.6	91.1	93.5	99.1
LoRA CLIP LC	32	64	68.3	74.9	78.8	90.9	86.5	92.2	94.7	99.2
LoRA CLIP LC	64	128	**69.4**	**75.4**	**79.7**	94.1	**88.2**	**93.2**	**96.0**	**99.5**

Additionally, we consider a traditional fine-tuned CLIP (FT CLIP LC) where we update all the weights of the backbone for the deepfake task. As highlighted in Table 4, a complete fine-tuning causes poor performance when $N \in \{20, 50, 100\}$. For instance, when training on ProGAN generator we report an accuracy of 52.6%, 54.8%, and 60.5% when training respectively on 20, 50, and 100 couples of real-fake images. This can be attributable to an overfitting scenario caused by a lack of training data. When complete fine-tuning is applied, the performance tends to increase when relaxing the few-shot constraint to $N = 1000$.

Validation on unseen generators. While the primary focus of this paper is on evaluating the efficacy of deepfake detectors in few-shot learning scenarios, our investigation extends to asses how the classifiers perform on images generated by generative models not encountered during their training phase. Specifically, we analyze results on different diffusion models, namely Guided [17], LDM [42], GLIDE [35] and an autoregressive generator in DALL-E [41]. Further, we analyze a selection of GAN-generated images from ProGAN [29], CycleGAN [58], BigGAN [6], StyleGAN [30], GauGAN [38], StarGAN [11], and other generative models, namely Deepfake [44], SITD [9], SAN [15], CRN [10], and IMLE [31].

We report in Table 5 and Table 6 the results of our LoRA CLIP and competitors on images generated by the previously mentioned generative models, following the datasets introduced by Ojha *et al.* [36] and Wang *et al.* [53] respectively. In addition to our baselines, we report the results obtained with the released checkpoints of the CLIP-based linear classifier introduced in [36] and both ResNet50 versions proposed in [53]. It is worth noting that while our proposal and baselines are trained on ProGAN with $N = 1000$, both the introduced competitors are trained on 360k real-fake pairs from ProGAN and LSUN.

Upon analysis of Table 5, LoRA CLIP exhibits superior performance over all baseline models, achieving accuracy improvements of 1.0%, 4.5%, and 11.7% in comparison to CLIP LC, SVM, and k-NN classifiers, respectively. These results underscore the effectiveness of LoRA-adapted embedding space in enhancing detection capabilities on unseen generators, towards a generalized deepfake detection embedding space. Compared to the CLIP linear classifier proposed in [36], our LoRA CLIP obtains comparable results with an average loss on performance of −0.4% but with leveraging 360 times fewer training samples.

Table 5. Accuracy results of detectors trained on ProGAN and tested on external generators [36] unseen during training. The symbol † represents pre-trained models, released by the authors, trained on 320k samples.

		LDM			GLIDE				
Model	Guided	200	200 (CFG)	100	100 (27)	50 (27)	100 (10)	DALL-E	**Avg**
CLIP LC† [36]	69.5	94.4	74	95.0	78.5	79.1	77.9	87.3	82.0
ResNet50 0.1† [53]	62.0	53.9	55.3	55.1	60.3	62.7	61.0	56.1	58.3
ResNet50 0.5† [53]	52.3	51.1	51.4	51.3	53.3	55.6	54.3	52.5	52.7
CLIP k-NN	61.3	73.6	67.2	73.9	71.4	72.1	71.3	68.8	69.9
CLIP SVM	63.5	85.4	64.5	87.3	82.0	82.0	81.8	70.2	77.1
CLIP LC	67.4	91.4	64.5	92.7	**87.1**	**86.1**	**85.5**	70.4	80.6
FT CLIP LC	54.3	76.4	67.1	77.1	61.9	62.1	62.9	72.9	66.8
LoRA CLIP LC	**68.4**	**93.9**	**68.3**	**94.4**	83.6	83.5	83.4	**77.3**	**81.6**

Differently, in Table 6 LoRA CLIP obtains the best result on average compared to all the analyzed methodologies. Specifically, our solution outperforms CLIP LC, SVM, and k-NN by respectively 9.0%, 10.2%, and 18.2% accuracy on average. Additionally, improvements of 4.7%, 5.3%, and 9.4% are obtained in comparison to the CLIP-based detector proposed in [36] and both detectors proposed by Wang *et al.* [53]. This further provides evidence of the efficacy of LoRA adaptation in the research field of deepfake detection. Furthermore, it is interesting to note that traditional fine-tuning (FT CLIP) loses generalization capabilities on unseen generators reporting a deficit accuracy on average of −14.8% and −8.3% when compared to LoRA CLIP in Table 5 and Table 6 respectively. This is likely due to the overfitting on the ProGAN generator observed during training. Fine-tuning all parameters completely modifies the deepfake subspace inside the CLIP embedding space, thus losing generalization capabilities.

Table 6. Accuracy results of detectors trained on ProGAN and tested on external generators [53] unseen during training. The symbol † represents pre-trained models, released by the authors, trained on 320k samples.

	Pro-GAN	Cycle-GAN	Big-GAN	Style-GAN	Gau-GAN	Star-GAN	Deep-Fake	SITD	SAN	CRN	IMLE	**Avg**
CLIP LC† [36]	99.8	98.3	95.1	84.9	99.5	95.8	68.6	62.2	56.6	56.6	69.1	80.6
ResNet50 0.1† [53]	100	85.2	70.2	87.1	78.9	91.8	53.5	90.3	50.5	86.3	86.2	80.0
ResNet50 0.5† [53]	100	80.8	59.0	73.4	79.3	81.0	51.1	78.3	50	87.6	94.1	75.9
CLIP k-NN	79.6	80.2	68.1	65.9	81.4	72.3	55.0	55.6	59.4	60.1	61.2	67.1
CLIP SVM	98.8	84.9	80.0	77.2	95.1	75.1	**63.4**	70.0	59.4	61.3	61.6	75.1
CLIP LC	99.1	85.9	81.3	77.8	96.9	69.8	61.2	71.1	64.2	61.4	70.2	76.3
FT CLIP LC	99.7	87.7	83.8	78.7	95.9	68.3	54.7	72.8	55.3	61.5	88.6	77.0
LoRA CLIP LC	**99.8**	**95.8**	**91.7**	**85.0**	**99.6**	**77.4**	59.2	**75.0**	**66.0**	**91.8**	**97.6**	**85.3**

5 Conclusion

In this study, we analyze the efficacy of CLIP-based deepfake detectors under conditions of few-shot learning, assessing their performance across various generators. Moreover, we introduce LoRA CLIP, aimed at refining the CLIP embedding space for the task of deepfake detection. The experimental results validate the effectiveness of our proposed method in identifying synthetic images within few-shot contexts. Further, the LoRA-enhanced CLIP model exhibits significant generalization capabilities to previously not encountered generative models.

Acknowledgments. We acknowledge the CINECA award under the ISCRA initiative, for the availability of high-performance computing resources and support. This work has been supported by the Horizon Europe project "European Lighthouse on Safe and Secure AI (ELSA)" (HORIZON-CL4-2021-HUMAN-01-03), co-funded by the European Union.

References

1. Amoroso, R., Morelli, D., Cornia, M., Baraldi, L., Del Bimbo, A., Cucchiara, R.: Parents and Children: Distinguishing Multimodal DeepFakes from Natural Images. ACM TOMM (2024)
2. Balaji, Y., Nah, S., Huang, X., Vahdat, A., Song, J., Zhang, Q., Kreis, K., Aittala, M., Aila, T., Laine, S., et al.: eDiff-I: Text-to-Image Diffusion Models with an Ensemble of Expert Denoisers. arXiv preprint arXiv:2211.01324 (2022)
3. Baraldi, L., Cocchi, F., Cornia, M., Baraldi, L., Nicolosi, A., Cucchiara, R.: Contrasting Deepfakes Diffusion via Contrastive Learning and Global-Local Similarities. In: ECCV (2024)
4. Betker, J., Goh, G., Jing, L., Brooks, T., Wang, J., Li, L., Ouyang, L., Zhuang, J., Lee, J., Guo, Y., et al.: Improving image generation with better captions (2023)
5. Betti, F., Staiano, J., Baraldi, L., Baraldi, L., Cucchiara, R., Sebe, N.: Let's ViCE! Mimicking Human Cognitive Behavior in Image Generation Evaluation. In: ACM Multimedia (2023)

6. Brock, A., Donahue, J., Simonyan, K.: Large Scale GAN Training for High Fidelity Natural Image Synthesis. In: ICLR (2018)
7. Bucciarelli, D., Moratelli, N., Cornia, M., Baraldi, L., Cucchiara, R., et al.: Personalizing Multimodal Large Language Models for Image Captioning: An Experimental Analysis. In: ECCV Workshops (2024)
8. Caffagni, D., Cocchi, F., Barsellotti, L., Moratelli, N., Sarto, S., Baraldi, L., Baraldi, L., Cornia, M., Cucchiara, R.: The Revolution of Multimodal Large Language Models: A Survey. In: ACL Findings (2024)
9. Chen, C., Chen, Q., Xu, J., Koltun, V.: Learning to See in the Dark. In: CVPR (2018)
10. Chen, Q., Koltun, V.: Photographic Image Synthesis with Cascaded Refinement Networks. In: ICCV (2017)
11. Choi, Y., Choi, M., Kim, M., Ha, J.W., Kim, S., Choo, J.: StarGAN: Unified Generative Adversarial Networks for Multi-Domain Image-to-Image Translation. In: CVPR (2018)
12. Cocchi, F., Baraldi, L., Poppi, S., Cornia, M., Baraldi, L., Cucchiara, R.: Unveiling the Impact of Image Transformations on Deepfake Detection: An Experimental Analysis. In: ICIAP (2023)
13. Corvi, R., Cozzolino, D., Poggi, G., Nagano, K., Verdoliva, L.: Intriguing Properties of Synthetic Images: From Generative Adversarial Networks to Diffusion Models. In: CVPR Workshops (2023)
14. Cozzolino, D., Poggi, G., Corvi, R., Nießner, M., Verdoliva, L.: Raising the Bar of AI-generated Image Detection with CLIP. In: CVPR Workshops (2024)
15. Dai, T., Cai, J., Zhang, Y., Xia, S.T., Zhang, L.: Second-Order Attention Network for Single Image Super-Resolution. In: CVPR (2019)
16. Dettmers, T., Pagnoni, A., Holtzman, A., Zettlemoyer, L.: QLoRA: Efficient Finetuning of Quantized LLMs. In: NeurIPS (2023)
17. Dhariwal, P., Nichol, A.: Diffusion Models Beat GANs on Image Synthesis. In: NeurIPS (2021)
18. Ding, M., Yang, Z., Hong, W., Zheng, W., Zhou, C., Yin, D., Lin, J., Zou, X., Shao, Z., Yang, H., et al.: CogView: Mastering Text-to-Image Generation via Transformers. In: NeurIPS (2021)
19. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. In: ICLR (2021)
20. Epstein, D.C., Jain, I., Wang, O., Zhang, R.: Online Detection of AI-Generated Images. In: ICCV Workshops (2023)
21. Esser, P., Rombach, R., Ommer, B.: Taming Transformers for High-Resolution Image Synthesis. In: CVPR (2021)
22. Frank, J., Eisenhofer, T., Schönherr, L., Fischer, A., Kolossa, D., Holz, T.: Leveraging frequency analysis for deep fake image recognition. In: ICML (2020)
23. Gadre, S.Y., Ilharco, G., Fang, A., Hayase, J., Smyrnis, G., Nguyen, T., Marten, R., Wortsman, M., Ghosh, D., Zhang, J., et al.: DataComp: In search of the next generation of multimodal datasets. In: NeurIPS (2024)
24. Grommelt, P., Weiss, L., Pfreundt, F.J., Keuper, J.: Fake or JPEG? Revealing Common Biases in Generated Image Detection Datasets. arXiv preprint arXiv:2403.17608 (2024)
25. He, K., Zhang, X., Ren, S., Sun, J.: Deep Residual Learning for Image Recognition. In: CVPR (2016)
26. Ho, J., Jain, A., Abbeel, P.: Denoising Diffusion Probabilistic Models. In: NeurIPS (2020)

27. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W.: LoRA: Low-Rank Adaptation of Large Language Models. In: ICLR (2022)
28. Hu, Z., Wang, L., Lan, Y., Xu, W., Lim, E.P., Bing, L., Xu, X., Poria, S., Lee, R.K.W.: LLM-Adapters: An Adapter Family for Parameter-Efficient Fine-Tuning of Large Language Models. arXiv preprint arXiv:2304.01933 (2023)
29. Karras, T., Aila, T., Laine, S., Lehtinen, J.: Progressive Growing of GANs for Improved Quality, Stability, and Variation. In: ICLR (2018)
30. Karras, T., Laine, S., Aila, T.: A Style-Based Generator Architecture for Generative Adversarial Networks. In: CVPR (2019)
31. Li, K., Zhang, T., Malik, J.: Diverse Image Synthesis from Semantic Layouts via Conditional IMLE. In: ICCV (2019)
32. Li, Y., Yang, X., Sun, P., Qi, H., Lyu, S.: Celeb-DF: A Large-scale Challenging Dataset for DeepFake Forensics. In: CVPR (2020)
33. Liao, W., Hu, K., Yang, M.Y., Rosenhahn, B.: Text to Image Generation with Semantic-Spatial Aware GAN. In: CVPR (2022)
34. Ni, H., Shi, C., Li, K., Huang, S.X., Min, M.R.: Conditional image-to-video generation with latent flow diffusion models. In: CVPR (2023)
35. Nichol, A., Dhariwal, P., Ramesh, A., Shyam, P., Mishkin, P., McGrew, B., Sutskever, I., Chen, M.: GLIDE: Towards Photorealistic Image Generation and Editing with Text-Guided Diffusion Models. In: ICML (2022)
36. Ojha, U., Li, Y., Lee, Y.J.: Towards Universal Fake Image Detectors That Generalize Across Generative Models. In: CVPR (2023)
37. Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., et al.: Training language models to follow instructions with human feedback. NeurIPS (2022)
38. Park, T., Liu, M.Y., Wang, T.C., Zhu, J.Y.: Semantic Image Synthesis with Spatially-Adaptive Normalization. In: CVPR (2019)
39. Poppi, S., Poppi, T., Cocchi, F., Cornia, M., Baraldi, L., Cucchiara, R.: Safe-CLIP: Removing NSFW Concepts from Vision-and-Language Models. In: ECCV (2024)
40. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning Transferable Visual Models From Natural Language Supervision. In: ICML (2021)
41. Ramesh, A., Pavlov, M., Goh, G., Gray, S., Voss, C., Radford, A., Chen, M., Sutskever, I.: Zero-shot text-to-image generation. In: ICML (2021)
42. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: CVPR (2022)
43. Ronneberger, O., Fischer, P., Brox, T.: U-Net: Convolutional Networks for Biomedical Image Segmentation. In: MICCAI (2015)
44. Rossler, A., Cozzolino, D., Verdoliva, L., Riess, C., Thies, J., Nießner, M.: Faceforensics++: Learning to detect manipulated facial images. In: ICCV (2019)
45. Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E.L., Ghasemipour, K., Gontijo Lopes, R., Karagol Ayan, B., Salimans, T., et al.: Photorealistic Text-to-Image Diffusion Models with Deep Language Understanding. In: NeurIPS (2022)
46. Schuhmann, C., Vencu, R., Beaumont, R., Kaczmarczyk, R., Mullis, C., Katta, A., Coombes, T., Jitsev, J., Komatsuzaki, A.: LAION-400M: Open Dataset of CLIP-Filtered 400 Million Image-Text Pairs. In: NeurIPS Workshops (2021)
47. Sha, Z., Li, Z., Yu, N., Zhang, Y.: DE-FAKE: Detection and Attribution of Fake Images Generated by Text-to-Image Generation Models. In: ACM CCS (2023)
48. Shah, V., Ruiz, N., Cole, F., Lu, E., Lazebnik, S., Li, Y., Jampani, V.: ZipLoRA: Any Subject in Any Style by Effectively Merging LoRAs. arXiv preprint arXiv:2311.13600 (2023)

49. Sohl-Dickstein, J., Weiss, E., Maheswaranathan, N., Ganguli, S.: Deep Unsupervised Learning using Nonequilibrium Thermodynamics. In: ICML (2015)
50. Tao, M., Bao, B.K., Tang, H., Xu, C.: GALIP: Generative Adversarial CLIPs for Text-to-Image Synthesis. In: CVPR (2023)
51. Wang, R., Juefei-Xu, F., Ma, L., Xie, X., Huang, Y., et al.: FakeSpotter: A Simple yet Robust Baseline for Spotting AI-Synthesized Fake Faces. In: IJCAI (2020)
52. Wang, S., Chen, L., Jiang, J., Xue, B., Kong, L., Wu, C.: LoRA Meets Dropout under a Unified Framework. arXiv preprint arXiv:2403.00812 (2024)
53. Wang, S.Y., Wang, O., Zhang, R., Owens, A., Efros, A.A.: CNN-generated images are surprisingly easy to spot...for now. In: CVPR (2020)
54. Wang, Z., Bao, J., Zhou, W., Wang, W., Hu, H., Chen, H., Li, H.: DIRE for Diffusion-Generated Image Detection. In: ICCV (2023)
55. Yang, X., Li, Y., Lyu, S.: Exposing deep fakes using inconsistent head poses. In: ICASSP (2019)
56. Yu, F., Zhang, Y., Song, S., Seff, A., Xiao, J.: LSUN: Construction of a Large-scale Image Dataset using Deep Learning with Humans in the Loop. arXiv preprint arXiv:1506.03365 (2015)
57. Yu, J., Xu, Y., Koh, J.Y., Luong, T., Baid, G., Wang, Z., Vasudevan, V., Ku, A., Yang, Y., Ayan, B.K., et al.: Scaling Autoregressive Models for Content-Rich Text-to-Image Generation. arXiv preprint arXiv:2206.10789 (2022)
58. Zhu, J.Y., Park, T., Isola, P., Efros, A.A.: Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks. In: ICCV (2017)

Multi-source Deep Domain Adaptation for Deepfake Detection

Md Shamim Seraj and Shayok Chakraborty(✉)

Department of Computer Science, Florida State University, Tallahassee, USA
shayok@cs.fsu.edu

Abstract. With the unprecedented success of generative models like GANs, synthetic image manipulations such as deepfakes have emerged as a serious concern in today's world. Existing techniques demonstrate promise in detecting deepfakes on which they are trained; however, their performance drops significantly when applied to detect forgeries created using other manipulation techniques, on which the model has not been sufficiently trained. Thus, detecting new types of deepfakes without losing prior knowledge about already learned faking techniques, is a problem of immense practical importance. In this paper, we propose a novel multi-source deep domain adaptation framework to address this challenge. Our framework can leverage a large amount of labeled data (fake/genuine) generated using one or more faking techniques (source domains) and a small amount of labeled data generated using a target faking technique of interest (target domain) to induce a deep neural network with good generalization capability on all the source domains, as well as the target domain. Further, our framework can efficiently utilize unlabeled data in the target domain, which is more readily available than labeled data. We design a novel loss function specific to the multi-source domain adaptation task and use the SGD method to optimize the loss and train the deep network. Our extensive empirical studies on benchmark datasets, using different types of deepfakes, corroborate the promise and potential of our framework for real-world applications. To the best of our knowledge, this is the first research effort to develop a multi-source deep domain adaptation technique for deepfake detection.

Keywords: Deepfake detection · Domain adaptation · Deep learning

1 Introduction

The inception of image manipulation dates back to photography itself [8]; however, with the recent advances in generative AI (and the advent of models such as generative adversarial networks or GANs [10] and autoencoders [11]), it has

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/978-3-031-78305-0_9.

A. Antonacopoulos et al. (Eds.): ICPR 2024, LNCS 15321, pp. 127–142, 2025.
https://doi.org/10.1007/978-3-031-78305-0_9

reached an unprecedented level of sophistication. Intense research in computer vision and machine learning has resulted in the generation of manipulation tools which are extremely easy to use and within everybody's reach. Of particular concern are human facial manipulations created using deep learning techniques, called *deepfakes* [19] [39] [47] [1]. For instance, using the *FaceSwap* deepfake, an attacker can put the victim in place and settings they have never been. It is also fairly easy to generate entirely synthetic faces even at very high resolutions [13], animate a subject's face to make it express the desired emotions [37], or modify facial expressions [45]. These techniques can be used for several malicious purposes, such as creating child sexual abuse materials, celebrity pornographic videos and fake propaganda videos for gaining unlawful political influence [51] [35] [7]. Against the backdrop of such growing concerns, deepfake detection has gained increasing research attention in the machine learning community [29]. Techniques using deep Convolutional Neural Networks (CNNs) have depicted promising performance for deepfake detection [1] [3] [21] [40] [54] [59] [60].

While effective, the deep neural networks (DNNs) often tend to overfit to the manipulation specific artifacts and learn feature representations accordingly. Thus, while they depict impressive performance on deepfakes on which the models have been trained, the learned features lack transferability, and their performance drops drastically when applied on new types of forgeries, even though they are semantically similar [56] [46] [6] [17]. Accurately detecting the new forgeries necessitates abundant labeled data (fake / genuine) from the new domain. However, the field of digital forensics is progressing at a rapid pace and researchers and practitioners are developing newer and sophisticated faking techniques on a regular basis (for instance, OpenAI recently released *Sora* [2], that can create realistic and imaginative scenes from text instructions). Thus, obtaining a large amount of labeled data for every single forgery technique is not feasible. Our goal is to detect forgeries given only a few (or none) labeled samples and a moderate amount of unlabeled samples from the new faking technique (unlabeled data is more readily obtained than labeled data; for instance, we may have access to several images which may or may not be forged using the new faking technique, and that information is not available to us, that is, the labels of these images are unknown). Further, while accurately detecting new types of deepfakes is important, it is also equally important to retain prior knowledge about already learned faking techniques, to avoid the *catastrophic forgetting* problem [14] [48] [53].

We thus pose the research task as follows: *we are given abundant data generated using different types of deepfakes (such as FaceSwap, Face2Face etc.); each type of deepfake constitutes a particular source domain data. The data in all the source domains are all labeled (genuine / fake). We are also given a small amount (or none) of labeled data and a moderate amount of unlabeled data generated using target faking technology of interest (target domain data). Our objective is to train a deep CNN to accurately identify fake and genuine images in all the source domains, as well as the target domain of interest.*

[1] we use *deepfake* as a generic term to denote any kind of image manipulation

[2] https://openai.com/sora

In this paper, we propose a novel multi-source deep domain adaptation technique to address this important problem in deepfake detection research. *Domain Adaptation (DA)* or *Transfer Learning (TL)* algorithms utilize abundant labeled data in one or more domains to develop a model for a related domain of interest, where labeled data is scarce [33]. The domain of interest is referred to as the *target* domain and all the other domains are called the *source* domains. The probability distributions generating the data in all the source and target domains are different, which implies that a deep model trained on the source domain data may not directly generalize to the target domain. Our contributions in this paper can be summarized as follows:
(*i*) We address the problem of multi-source DA for deepfake detection. We propose a novel loss function specifically designed for the multi-source DA task and utilize the SGD algorithm to optimize the loss and train a deep CNN. To our knowledge, multi-source domain adaptation has not been studied in the context of deepfake detection.
(*ii*) We propose a strategy to leverage unlabeled data in the target domain in training the deep CNN, which is more readily available than labeled data.
(*iii*) We extensively validate our framework on benchmark datasets with a variety of faking techniques, on challenging low resolution data, and also with varied number of labeled images from the target domain. Our framework depicts impressive performance even when the target domain contains only unlabeled samples, and no labeled data is available in the target domain.

2 Related Work

Deepfake Detection: With the advent of sophisticated faking techniques, there has been an increasing interest in developing deepfake detection technology in the research community. Most of the current detection techniques use deep neural networks (DNNs) [28] [46] [38]. These methods attempt to detect specific artifacts in the data, such as abnormal eye blinking [22], signal level artifacts [23], irregular head poses [54] and peculiar behavior patterns [2] among others. As mentioned before, these methods suffer from poor generalization, when tested on deepfakes of a different type than those in the training data.

Domain Adaptation: Domain Adaptation (DA) algorithms transfer relevant knowledge from a source domain with abundant labeled data, to a target domain of interest, where labeled data is scarce, under the constraint of a probability distribution difference between the domains [33]. DA techniques using deep learning have outperformed DA using hand-engineered features [32] [34]. Several metrics have been studied to quantify the disparity between the source and target domains and learn domain invariant features using a DNN, such as the Maximum Mean Discrepancy (MMD) [49] [27] [50], the Kullback Leibler Divergence [31], the Jensen Shannon Divergence [43] among others. Techniques based on adversarial training has depicted particularly impressive performance in DA; algorithms in this category include the Domain Adversarial Neural Network (DANN) [9], the Coupled Generative Adversarial Network (CoGAN) and its combination with Variational Autoencoder (VAE) [18], and Wasserstein GAN [42] among others. Multi-source domain adaptation is an extension of DA, where data from multiple source domains are available. Peng *et al.* [36] proposed the moment matching algorithm for multi-source DA, which aims to transfer knowledge learned from multiple labeled source domains to an unlabeled target domain by dynamically aligning moments of their feature distributions. Adversarial training techniques are also popular in aligning multiple source domains and the target domain [57] [52] (similar to single source DA).

Domain Adaptation for Deepfake Detection: Even though both DA and deepfake detection have been extensively studied, DA for deepfake detection is much less explored. Kim *et al.* [17] employed representation learning and knowledge distillation to introduce *FReTAL*, a transfer learning-based method for deepfake detection. They also proposed the *CoReD* framework by combining the ideas of continual learning, representation learning and knowledge distillation to perform sequential DA on new deepfake datasets [16]. Tariq *et al.* [46] adopted a fine-tuning strategy for DA, where the DNN was first trained on the source domain data and then the deeper layers were fine-tuned using the target domain data. Fine-tuning was also adopted by Cozzolino *et al.* [6] in their *ForensicTransfer* framework, which first learned an autoencoder on the source domain data, which was then fine-tuned using target domain data. All these methods share a common drawback: they all require all the samples in the target domain to be labeled. As mentioned in Section 1, we may encounter a situation where we have access to a large number of images, but we are unable to verify whether they are forged using the new faking mechanism of interest or not, that is, the labels of these images are not available to us. Very recently, researchers have begun to explore DA techniques which can utilize unlabeled data in the target domain. Chen and Tan [5] and Seraj *et al.* [41] used the concept of adversarial DA for deepfake detection. Zhang *et al.* [56] trained a CNN to learn domain invariant features by minimizing the MMD between the source and target domains. These methods require access to the domain labels (whether a sample is derived from the source or target domain), rather than the task labels (fake/genuine) and can thus leverage unlabeled data in the target domain. However, all the aforementioned methods assume a single source and a single target domain, and do not address the challenge of detecting deepfakes in the target domain, when data from multiple source domains are available.

In our framework, we use the concept of moment matching to address the disparity between multiple source domains and the target domain. We also formulate a class alignment loss term on the unlabeled target domain data, which enforces each target sample to align to exactly one of the source classes (genuine/fake) and be distinct from the other class. We conduct extensive empirical studies on benchmark datasets with several types of deepfakes, under challenging real-world conditions, such as low-resolution images (which are common in social media) and very few (including none) labeled samples from the target domain of interest. We now describe our framework.

3 Proposed Framework

3.1 Problem Setup

We are given data from N_S source domains $S_1, S_2, \dots S_{N_S}$ and a target domain T. Each domain represents data generated using a particular forgery technique (such as *FaceSwap*, *Face2Face* etc.). Due to the difference in the faking technologies, the data in each domain is derived from a different probability distribution. The data in all the source domains are all labeled: $S_i = \{x_j, y_j\}_{j=1}^{|S_i|}, \forall i = 1, \dots N_S$. Here $\{x_j\}$ denotes the deep feature representation of a particular image and $\{y_j\}$ denotes the binary label (fake/genuine). Let D_S denote the combined data from all the source domains: $D_S = S_1 \cup S_2 \cup \dots \cup S_{N_S}$. We are also given data from a target domain of interest, which contains labeled samples: $D_T^L = \{x_j, y_j\}_{j=1}^{N_T^L}$, as well as unlabeled samples: $D_T^U = \{x_j\}_{j=1}^{N_T^U}$. As explained in Section 1, we do not have sufficient supervision in the target domain, that is, $|D_T^L| \ll |D_T^U|$. Let $D_T = D_T^L \cup D_T^U$. Our objective is to train a deep convolutional neural network (CNN) which will furnish good generalization performance on all the source domains, as well as the target domain; that is, we would like our trained CNN to reliably detect deepfakes generated using all the faking techniques. We propose to formulate a novel loss function, specific to the task, and train the deep network to optimize that loss. Our loss function consists of three components: (i) supervised loss on the labeled data, which encourages the network to incur minimal prediction error on the labeled source and labeled target samples; (ii) a moment matching loss to align the source and the target data distributions, and learn domain invariant feature representations accordingly; and (iii) unsupervised loss on unlabeled target data, which encourages the network to predict the unlabeled target samples with high confidence. These are detailed in the following sections.

3.2 Supervised Loss on the Labeled Source and Labeled Target Data

The goal of the supervised loss term is to ensure that the network learns feature representations to accurately classify the labeled source and target samples. Let D_L denote the labeled source and target samples: $D_L = D_S \cup D_T^L = \{x_1, x_2, \dots, x_{n_L}\}$, with corresponding labels $\{y_1, y_2, \dots, y_{n_L}\}$. Since the labels are binary in our problem (fake/genuine), we use the binary cross entropy (BCE) [26] as the supervised loss to train the deep CNN:

$$\mathcal{L}_{sup} = -\frac{1}{n_L} \sum_{i=1}^{n_L} \Big(y_i . \log(p(y_i)) + (1 - y_i) . \log(1 - p(y_i)) \Big) \tag{1}$$

where $p(y_i)$ denotes the probability obtained from the softmax activation layer of the CNN.

3.3 Moment Matching Loss on Source and Target Data

The moments of distributions have been studied by the machine learning community for a variety of applications. In the context of domain adaptation, the maximum mean discrepancy (MMD) metric has been used in previous research, which aligns the first order moments of two distributions [27] [50]. Matching second order and higher order moments have also been studied for DA [44] [55]. With the advent of GANs, many GAN-based moment matching approaches have been proposed, such as McGAN [30], GMMN [25] and MMD GAN [20]. Our strategy to address the disparity between the source and target domains is motivated by the moment matching algorithm proposed by Peng *et al.* [36]. This method not only aligns all the source domains with the target domain, but also aligns the source domains with each other simultaneously, by directly aligning the moments of their deep feature distributions. The loss function to align the moments of the source and target distributions can be expressed as follows:

$$\mathcal{L}_{MM} = \sum_{m=1}^{M} \left(\frac{1}{N_S} \sum_{i=1}^{N_S} \|\mathbb{E}(X_i^m) - \mathbb{E}(X_T^m)\|_2 + \frac{1}{\binom{N_S}{2}} \sum_{i,j=1, i \neq j}^{N_S} \|\mathbb{E}(X_i^m) - \mathbb{E}(X_j^m)\|_2 \right) \quad (2)$$

Here, X denotes the deep feature representations of the images obtained from the underlying deep neural network, $\mathbb{E}(.)$ denotes the expectation operator and $\|.\|_2$ denotes the vector 2-norm. The first term inside the parentheses attempts to align all the source distributions with the target distribution (by minimizing the distance between their moments), while the second term aligns the moments of the source distributions among themselves (considering one pair at a time). M is the maximum order of the moments considered; we used $M = 2$ in our empirical studies. Minimizing this term ensures that all the source and the target domains are aligned, and the CNN learns domain invariant feature representations. Note that, while our domain alignment strategy is motivated by the moment matching technique of Peng *et al.* [36], we did not use their network architecture or the other specific components in their pipeline (such as the feature extractor, source domain classifiers etc.). We further formulated a loss term (described next) to utilize the unlabeled target domain data in training the deep network, which was not done in [36].

3.4 Unsupervised Loss on Unlabeled Target Data

As mentioned in Section 1, our framework can leverage the presence of unlabeled data in the target domain of interest, to train the CNN to learn better feature representations. From a practical standpoint, we may not have access to a sufficient amount of labeled samples from the new deepfake that we are trying to detect; but we may have access to a large number of images which may or may not have been forged using the new technique (unlabeled data). We formulate a loss term to utilize the presence of such unlabeled data in the target domain. Our method is inspired by research in semi-supervised learning, where feature representations are learned such that the trained CNN furnishes confident predictions on the unlabeled samples [4] [50] [41]. Each unlabeled target domain sample can belong to one of the two classes: *fake (1)* and *genuine (2)*. We assume the presence of K samples from each class j in the labeled source data D_S, where $j \in \{1, 2\}$. Let $\mathcal{F}_S^{jk}$ denote the learned feature representation of the k^{th} source sample from class j and $\mathcal{F}_T^i$ denote the feature representation of an unlabeled target sample x_i. The fundamental rationale is to ensure that $\mathcal{F}_T^i$ is similar to all the K learned source representations from one of the classes j, and dissimilar to the other class. We enforce the similarity with K source samples (instead of a single sample) to result in a more robust alignment of the target data sample. We define a measure to capture the extent of the alignment, which quantifies the probability that the target sample x_i is assigned to class j:

$$p_{ij} = \frac{\sum_{k=1}^{K} exp\langle \mathcal{F}_T^i, \mathcal{F}_S^{jk} \rangle}{\sum_{j=1}^{2} \sum_{k=1}^{K} exp\langle \mathcal{F}_T^i, \mathcal{F}_S^{jk} \rangle} \tag{3}$$

Here, $\langle \cdot, \cdot \rangle$ denotes the dot product between two vectors (used to compute their similarity), the exponential function $exp(.)$ has been used for ease of differentiability and the denominator ensures that the measure is normalized, that is, $\sum_j p_{ij} = 1$. Ideally, for a given unlabeled target sample x_i, we would want one of the probabilities $p_{ij}, j \in \{1, 2\}$, to be high and the other to be low, denoting that the target sample is similar to exactly one of the source classes (fake/genuine) and dissimilar to the other class. In that case, p_i tends to be a one-hot vector, which can be interpreted as the model having low prediction uncertainty (entropy). We therefore define the unsupervised loss term as the entropy of the target probability vectors:

$$\mathcal{L}_{unsup} = -\frac{1}{N_T^U} \sum_{i=1}^{N_T^U} \sum_{j=1}^{2} p_{ij} \log p_{ij} \tag{4}$$

where N_T^U denotes the number of unlabeled target samples. Minimizing this loss produces probability vectors p_i that tend to be one-hot vectors, that is, the deep network furnishes confident predictions on the unlabeled target data. Note that the probability values in Equation (4) are derived using the class alignment score in Equation (3) and not using class prediction probabilities, as done conventionally. The overall loss function to train the deep CNN can thus be expressed as:

$$\mathcal{L} = \mathcal{L}_{sup} + \lambda_1 \mathcal{L}_{MM} + \lambda_2 \mathcal{L}_{unsup} \tag{5}$$

where λ_1 and λ_2 are weight parameters governing the relative importance of the terms. We use the stochastic gradient descent (SGD) method to optimize the loss and train the deep network.

4 Experiments and Results

Datasets: We used the **FaceForensic++ (FF++)** benchmark dataset [38] in our experimental studies. It contains $1,000$ pristine videos and $1,000$ fake videos generated using each of the following four faking techniques: *Face2Face (F2F), FaceSwap (FS), DeepFakes (DF)* and *NeuralTextures (NT)*. Each video was split to generate 50 images (128×128) and we used an off-the-shelf face recognition software [3] to detect and crop the facial regions from these images. We also conducted a cross-dataset study, where we used the Celeb-DF dataset [24], which contains high-quality deepfake videos of celebrities .

Comparison Baselines: We used DA algorithms that have been explicitly studied for the deepfake detection problem as comparison baselines in our work: (i) *Fine-tuning (FT)* [46]; (ii) *Transferable GAN-images Detection (TGD)* [12]; (iii) *Feature Representation Transfer Adaptation Learning (FReTAL)* [17]; (iv) *Knowledge Distillation (KD)* [17]; and (v) *FeatureTransfer (FeatTran)* [5]. All the baselines, except *FeatTran* are supervised, that is, they require all the samples in the target domain to be labeled (*FeatTran* was preferred over the MMD-based unsupervised DA method for deepfake detection [56], as it uses GAN-based adversarial domain alignment which is more popular in the DA community than MMD based alignment). Further, all these baselines are only designed for a single source and a single target domain; to extend them to the multi-source setting, we combined all the source domain data and treated that as a single source domain. This is the most straightforward strategy to extend single source DA to the multi-source setting and has been used in previous research on multi-source DA [36].

Experimental Setup: In each experiment, we were given images from multiple source domains and a target domain, where each domain represents a particular faking technique (*DF, F2F, FS* etc.). We experimented with two and three source domains in this research. The data in all the source domains were all labeled. The target domain contained a small amount of labeled data and a large amount of unlabeled data (to appropriately capture a real-world situation). We used $40,000$ images as the source domain data (taken equally from all the source domains), $2,000$ images as the labeled target domain data and $18,000$ images as the unlabeled target domain data. The test set contained $10,000$ images taken equally from all the source domains and $10,000$ images from the target domain, to validate the performance of the model on all the domains. Each experiment was conducted 3 times, and the results were averaged to rule out the effects of randomness. The parameters λ_1 and λ_2 were both taken as 1 in our experiments.

We used the *Xception* network as the base model due to its promising performance on the FaceForensics++ dataset [39]. A schematic diagram of the network architecture is shown in the Appendix. The F1 score was used as the evaluation metric on the test set, similar to [17].

[3] https://pypi.org/project/face-recognition/

Implementation Details: Please refer to the Appendix for details about the implementation. Our code will be made publicly available upon acceptance of our paper.

4.1 Main Results

Table 1 reports the results on 6 multi-source DA tasks (the notation $x, y, z \rightarrow w$ implies that x, y, z are the source domains and w is the target domain). The supervised DA techniques *(FReTAL, KD, FT and TGD)* cannot utilize the unlabeled samples in the target domain, and thus depict much lower accuracy. *FeatTran* utilizes the unlabeled target domain samples merely for domain alignment and does not involve any other strategy to use the information in the unlabeled target data. Thus, although it depicts better accuracy than the supervised methods (in most cases), its performance is not as good as the proposed framework. Further, none of these techniques are designed to handle multiple source domains, and combining all the source domains into one results in sub-optimal performance. Our framework can efficiently utilize unlabeled target domain data in training the deep model through the class alignment loss term; it can also address the disparity among multiple source domains and the target domain and efficiently leverage data from all the domains. It thus consistently depicts impressive performance and comprehensively outperforms all the baselines, for all the 6 tasks (involving two and three source domains). Our framework not only depicts high accuracy in identifying deepfakes in the new target domain, but also retains already learned knowledge from all the source domains. The performance improvement achieved by our method is quite substantial in some cases; for instance, for the *DF,FS,NT* → *sF2F* experiment, the performance improvement achieved by the proposed framework on the target domain is almost 10% compared to the closest baseline (*FReTAL*). These results unanimously corroborate the promise and potential of our method for out-of-domain deepfake detection in real-world applications.

Table 1. Mean (± std) F1 scores (in percentage) of all the methods for 6 out-of-domain deepfake detection tasks. Best F1 values are marked in **bold**. The notation $x, y, z \rightarrow w$ implies that x, y, z are the source domains and w is the target domain. Results are averaged over 3 runs.

DA Task	**Domain**	**Proposed**	**FeatTran** [5]	**FReTAL** [17]	**KD** [17]	**FT** [46]	**TGD** [12]
DF,FS → F2F	Source	**98.34** ± 0.75	94.33 ± 0.22	88.6 ± 1.08	93.76 ± 0.20	87.15 ± 0.69	88.9 ± 1.46
	Target	**96.46** ± 0.62	87.22 ± 0.06	85.08 ± 0.42	70.98 ± 3.52	81.88 ± 0.08	76.16 ± 4.29
FS,F2F → DF	Source	**96.54** ± 2.20	94.14 ± 0.12	85.37 ± 4.01	91.78 ± 1.61	80.26 ± 1.76	87.02 ± 3.00
	Target	**96.72** ± 1.37	89.80 ± 0.36	89.61 ± 0.92	84.13 ± 0.81	90.79 ± 0.13	88.16 ± 1.26
DF,F2F → FS	Source	**96.15** ± 0.52	94.60 ± 0.05	86.98 ± 1.62	92.74 ± 1.30	80.84 ± 1.28	87.84 ± 0.33
	Target	**96.05** ± 0.41	87.33 ± 0.09	87.40 ± 0.89	75.45 ± 3.38	86.47 ± 0.04	80.47 ± 2.39
FS,F2F,NT → DF	Source	**96.96** ± 0.87	88.34 ± 0.01	84.31 ± 1.18	84.91 ± 1.41	74.62 ± 0.76	80.84 ± 0.78
	Target	**96.75** ± 0.22	86.88 ± 0.17	89.07 ± 0.31	88.07 ± 0.18	90.21 ± 0.49	89.08 ± 0.29
DF,F2F,NT → FS	Source	**97.12** ± 1.15	89.55 ± 0.15	70.94 ± 2.81	83.75 ± 4.31	70.41 ± 1.69	78.11 ± 0.57
	Target	**97.76** ± 0.90	82.96 ± 0.20	85.39 ± 0.30	71.56 ± 0.92	83.37 ± 0.67	79.17 ± 1.30
DF,FS,NT → F2F	Source	**97.62** ± 0.66	88.82 ± 0.09	84.35 ± 2.00	88.36 ± 1.01	80.72 ± 0.34	83.97 ± 2.27
	Target	**96.06** ± 1.45	85.09 ± 0.53	86.32 ± 0.46	80.25 ± 2.19	83.91 ± 1.07	77.74 ± 4.77

4.2 Performance in Detecting Low Resolution Deepfakes

Data in social media often have low quality and resolutions [16]. We validate the performance of our framework in identifying low quality deepfakes in this experiment. We used images of 64 × 64 resolution for this experiment (as opposed to 128 × 128 used in the previous experiment). The results are presented in Table 2. Our framework once again outperforms all the baselines and achieves the highest F1 score in the source and target domains consistently across all the tasks. Here also, the performance improvement is quite substantial; for the *DF,FS* → *F2F* experiment, the highest F1 score achieved by the baselines on the target domain is 82.35, while our framework achieves an F1 score of 95.14, demonstrating an improvement of almost 13%. This shows the robustness of our framework to accurately identify deepfakes even in low quality data, and its usefulness in detecting deepfakes in social media.

Table 2. Mean (± std) F1 scores (in percentage) of all the methods for detecting low resolution deepfakes. Best F1 values are marked in **bold**. The notation $x, y, z \rightarrow w$ implies that x, y, z are the source domains and w is the target domain. Results are averaged over 3 runs.

DA Task	**Domain**	**Proposed**	**FeatTran** [5]	**FReTAL** [17]	**KD** [17]	**FT** [46]	**TGD** [12]
DF,FS → F2F	Source	**97.55** ± 1.02	92.38 ± 0.11	87.69 ± 1.33	90.37 ± 3.19	86.94 ± 1.68	85.14 ± 1.56
	Target	**95.14** ± 2.10	82.35 ± 0.54	80.65 ± 1.84	64.82 ± 2.61	70.88 ± 6.66	79.47 ± 0.43
FS,F2F → DF	Source	**96.03** ± 1.71	91.94 ± 0.09	89.32 ± 0.46	91.25 ± 1.07	86.00 ± 2.74	83.95 ± 0.39
	Target	**95.34** ± 1.36	88.73 ± 0.23	88.97 ± 0.51	81.53 ± 3.38	85.61 ± 3.05	90.66 ± 0.10
FS,F2F,NT → DF	Source	**94.45** ± 3.85	84.61 ± 0.18	81.12 ± 0.95	81.00 ± 2.28	79.49 ± 4.08	73.06 ± 1.28
	Target	**94.27** ± 3.85	83.68 ± 0.10	88.49 ± 0.43	85.65 ± 0.62	88.07 ± 0.78	90.28 ± 0.68
DF,FS,NT → F2F	Source	**96.79** ± 1.61	85.74 ± 0.45	83.31 ± 5.70	85.87 ± 0.78	82.66 ± 1.27	79.43 ± 0.89
	Target	**95.77** ± 1.51	80.58 ± 0.03	78.54 ± 6.40	79.96 ± 0.34	79.42 ± 0.91	82.27 ± 0.69

4.3 Cross Dataset Study

We explored a different scenario in this study, where the source domain data is derived from one dataset and the target domain data from a different dataset. Our goal was to study whether the knowledge of deepfakes from one dataset can be transferred to detecting image manipulations from a different dataset. We conducted two experiments, one where the source domain data were taken from FF++ and the target domain data from Celeb-DF [24]; and another, where the source domain data were from FF++ and the target domain data from DFDC [4]. The images in Celeb-DF and DFDC are manipulated using deepfake synthesis algorithms, different from that used in FF++. The results are reported in Table 3 for the Celeb-DF dataset and Table 4 for the DFDC dataset. Consistent with the previous results, our method achieves the highest F1 scores on both the source and target domains for both the experiments, for both the multi-source DA tasks. This shows that our framework is not only able to transfer relevant knowledge from FF++ to accurately detect new types of deepfakes in Celeb-DF and DFDC,

[4] https://ai.meta.com/datasets/dfdc/

but also retains its knowledge about the deepfakes in FF++, and thus effectively mitigates catastrophic forgetting. These results further corroborate the practical usefulness of our framework.

4.4 Feature Visualizations

The goal of this experiment was to visualize the features learned by our framework through t-SNE embeddings. We used the *FeatTran* method as a comparison baseline, as it also utilizes unlabeled data in the target domain to train the deep CNN. Figure 1 depicts the results of four different deepfake detection tasks. For ease of interpretation, the data from all the source domains are represented with a *plus* sign, while the data from the target domain are represented with a *circle*; blue color denotes the *Real / Genuine* class and red color denotes the *Fake* class. As evident visually, the proposed method shows a better separation of the two categories (blue and red clusters) and a better overlap between the source and target domains, compared to *FeatTran*. Thus, using the moment matching and the unsupervised class alignment loss terms in our framework, the deep CNN is able to learn discriminating feature representations that minimize the disparity among all the source and target domains, and also separate the real and fake images from the two domains. Thus, it is able to achieve impressive F1 scores consistently across all the experiments.

Table 3. Mean (± std) F1 scores (in percentage) of all the methods for the cross dataset study on the Celeb-DF dataset. Best F1 values are marked in **bold**. The notation $x, y, z \rightarrow w$ implies that x, y, z are the source domains and w is the target domain. Results are averaged over 3 runs.

DA Task	**Domain**	**Proposed**	**FeatTran** [5]	**FReTAL** [17]	**KD** [17]	**FT** [46]	**TGD** [12]
FS,F2F → Celeb-DF	Source	**97.68** ± 1.31	94.47 ± 0.08	83.67 ± 10.47	69.94 ± 1.00	64.26 ± 1.55	63.42 ± 0.83
	Target	**94.91** ± 3.71	74.7 ± 0.82	73.22 ± 12.76	76.68 ± 0.05	86.62 ± 0.16	79.12 ± 1.91
FS,NT,F2F → Celeb-DF	Source	**97.14** ± 0.51	88.24 ± 0.41	90.30 ± 0.06	67.12 ± 0.37	60.21 ± 0.07	61.42 ± 6.56
	Target	**96.96** ± 0.74	73.70 ± 0.06	62.39 ± 0.33	75.06 ± 0.94	84.76 ± 0.12	74.75 ± 1.60

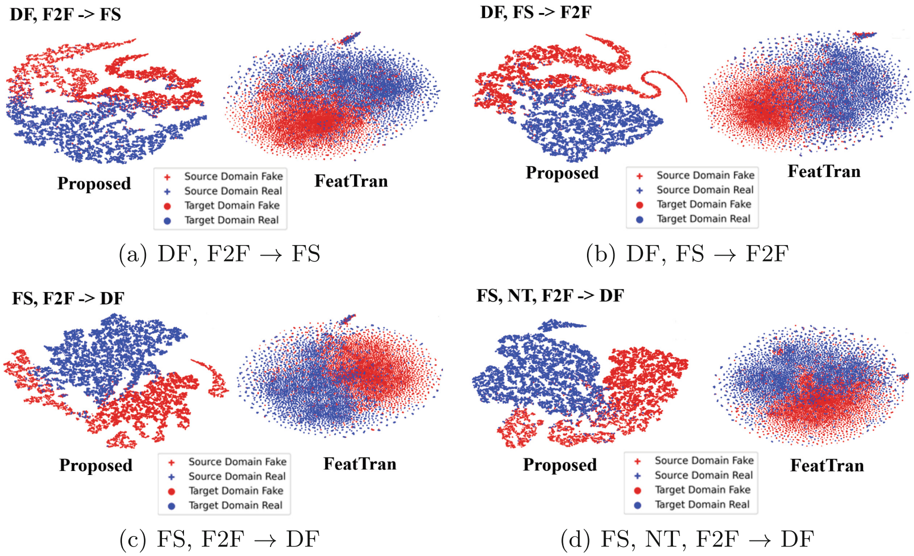
Table 4. Results for the cross dataset study on the DFDC dataset. Best F1 values are marked in **bold**. The notation $x, y \rightarrow w$ implies that x, y are the source domains and w is the target domain.

TL Task	**Domain**	**Proposed**	**FeatTran** [5]	**FReTAL** [17]	**KD** [17]	**FT** [46]	**TGD** [12]
DF,FS → DFDC	Source	**98.18**	96.53	82.31	75.29	65.45	75.70
	Target	**92.84**	91.18	90.63	78.33	90.55	89.38
FS,F2F → DFDC	Source	**98.24**	95.47	84.12	77.85	75.11	65.01
	Target	**92.42**	90.56	89.43	84.19	88.04	90.52

4.5 Ablation Study

We conducted ablation studies to assess the effects of the moment matching loss term $\mathcal{L}_{MM}$ and the unspervised loss term $\mathcal{L}_{unsup}$ in our framework (in Equation (5)). The F1 score on the target test set for two multi-source experiments are reported in Table 5. We note that the performance of our framework drops in the absence of the moment matching loss term; this shows the utility of this term to appropriately address the disparity among all the source domains and the target domain, and learn domain invariant features. The performance of our framework is also affected in the absence of the unsupervised class alignment loss term; this shows its usefulness to leverage the information in the unlabeled target domain data, and learn discriminating feature representations to improve the detection accuracy of our method.

We also conducted experiments to study the effects of the size of the labeled target domain data and the unlabeled target domain data on the detection performance. These results are included in the Appendix due to space constraints.



(a) DF, F2F → FS

(b) DF, FS → F2F

(c) FS, F2F → DF

(d) FS, NT, F2F → DF

Fig. 1. t-SNE visualization results. For ease of interpretation, the data from all the source domains are represented with a *plus* sign. The data from the target domain are represented with a *circle*. Blue color denotes the *Real / Genuine* class; red color denotes the *Fake* class. Best viewed in color.

Table 5. Ablation study results

DA Task	Proposed	Proposed w/o $\mathcal{L}_{unsup}$	Proposed w/o $\mathcal{L}_{MM}$
DF,FS → F2F	96.46	89.79	94.32
DF,F2F → FS	96.05	84.91	94.50

5 Conclusion and Future Work

With the tremendous progress of generative AI and the availability of sophisticated tools, deepfakes have become ubiquitous. While state-of-the-art CNNs have demonstrated promise in detecting deepfakes, their performance drops drastically when applied on out-of-domain data (deepfakes generated using a new faking technology). We proposed a novel multi-source DA technique using deep learning to address this challenging and practical issue. Contrary to existing methods, our framework can collate knowledge from multiple source domains and also utilize unlabeled target domain data efficiently for model training. Our extensive empirical studies demonstrated the promise of this method under challenging real-world conditions, such as low quality deepfakes and absence of labeled training data in the target domain of interest. To our knowledge, this research is the first of its kind to study the performance of multi-source deep domain adaptation techniques for deepfake detection. As part of future research, we plan to study the performance of our framework on other types of deepfakes besides images, such as audio [15] and text [58].

Acknowledgment. This research was supported in part by the National Science Foundation under Grant Number: IIS-2143424 (NSF CAREER Award).

References

1. Afchar, D., Nozick, V., Yamagishi, J., Echizen, I.: Mesonet: a compact facial video forgery detection network. In: IEEE International Workshop on Information Forensics and Security (WIFS). pp. 1–7 (2018)
2. Agarwal, S., Farid, H., El-Gaaly, T., Lim, S.: Detecting deep-fake videos from appearance and behavior. In: IEEE International Workshop on Information Forensics and Security (WIFS). pp. 1–6 (2020)
3. Agarwal, S., Farid, H., Gu, Y., He, M., Nagano, K., Li, H.: Protecting world leaders against deep fakes. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW) (2019)
4. Chapelle, O., Scholkopf, B., Zien, A.: Semi-Supervised Learning. The MIT Press (2006)
5. Chen, B., Tan, S.: Featuretransfer: Unsupervised domain adaptation for cross-domain deepfake detection. Security and Communication Networks (2021)
6. Cozzolino, D., Thies, J., Rossler, A., Riess, C., Niebner, M., Verdoliva, L.: Forensictransfer: Weakly-supervised domain adaptation for forgery detection. In: arXiv:1812.02510v2 (2019)
7. DelViscio, J.: A nixon deepfake, a ”moon disaster” speech and an information ecosystem at risk. Scientific American **20** (2020)
8. Farid, H.: Photo Forensics. The MIT Press (2016)
9. Ganin, Y., Ustinova, E., Ajakan, H., Germain, P., Larochelle, H., Laviolette, F., Marchand, M., Lempitsky, V.: Domain-adversarial training of neural networks. Journal of Machine Learning Research (JMLR) **17** (2016)
10. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. In: Neural Information Processing Systems (NeurIPS) (2014)

11. Goodfellow, I., Bengio, Y., Courville, A.: Deep Learning. MIT Press (2016)
12. Jeon, H., Bang, Y., Kim, J., Woo, S.S.: TGD: Transferable GAN-generated images detection framework. In: arXiv:2008.04115 (2020)
13. Karras, T., Laine, S., Aila, T.: A style-based generator architecture for generative adversarial networks. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2019)
14. Kemker, R., McClure, M., Abitino, A., Hayes, T., Kanan, C.: Measuring catastrophic forgetting in neural networks. In: AAAI Conference on Artificial Intelligence (2018)
15. Khanjani, Z., Watson, G., Janeja, V.: Audio deepfakes: A survey. Frontiers in Big Data **5** (2022)
16. Kim, M., Tariq, S., Woo, S.: Cored: Generalizing fake media detection with continual representation using distillation. In: ACM International Conference on Multimedia. pp. 337–346 (2021)
17. Kim, M., Tariq, S., Woo, S.S.: Fretal: Generalizing deepfake detection using knowledge distillation and representation learning. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). pp. 1001–1012 (2021)
18. Kingma, D., Welling, M.: Auto-encoding variational bayes. In: arXiv preprint arXiv:1312.6114 (2013)
19. Kowalski, M.: Faceswap - github repository. https://github.com/MarekKowalski/FaceSwap (2016), accessed: March 14, 2024
20. Li, C., Chang, W., Cheng, Y., Yang, Y., Poczos, B.: MMD GAN: Towards deeper understanding of moment matching network. In: Advances of Neural Information Processing Systems (NeurIPS) (2017)
21. Li, L., Bao, J., Zhang, T., Yang, H., Chen, D., Wen, F., Guo, B.: Face X-Ray for more general face forgery detection. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2020)
22. Li, Y., Chang, M., Lyu, S.: In Ictu Oculi: Exposing AI created fake videos by detecting eye blinking. In: IEEE International Workshop on Information Forensics and Security (WIFS). pp. 1–7. IEEE (2018)
23. Li, Y., Lyu, S.: Exposing deepfake videos by detecting face warping artifacts. In: arXiv:1811.00656 (2018)
24. Li, Y., Sun, P., Qi, H., Lyu, S.: Celeb-DF: A large-scale challenging dataset for deepfake forensics. In: IEEE Conference on Computer Vision and Patten Recognition (CVPR) (2020)
25. Li, Y., Swersky, K., Zemel, R.: Generative moment matching networks. In: International Conference on Machine Learning (ICML) (2015)
26. Liu, J., Chang, W., Wu, Y., Yang, Y.: Deep learning for extreme multi-label text classification. In: ACM SIGIR Conference on Information Retrieval (2017)
27. Long, M., Cao, Y., Wang, J., Jordan, M.: Learning transferable features with deep adaptation networks. In: International Conference on Machine Learning (ICML) (2015)
28. Malik, A., Kuribayashi, M., Abdullahi, S., Khan, A.: Deepfake detection for human face images and videos: A survey. IEEE Access **10** (2022)
29. Mirsky, Y., Lee, W.: The creation and detection of deepfakes: A survey. ACM Computing Surveys (CSUR) **54**(1), 1–41 (2021)
30. Mroueh, Y., Sercu, T., Goel, V.: McGAN: Mean and covariance feature matching GAN. In: International Conference on Machine Learning (ICML) (2017)
31. Nguyen, A., Tran, T., Gal, Y., Torr, P., Baydin, A.: Kl guided domain adaptation. In: International Conference on Learning Representations (ICLR) (2022)

32. Pan, S., Tsang, I., Kwok, J., Yang, Q.: Domain adaptation via transfer component analysis. In: International Joint Conference on Artificial Intelligence (IJCAI) (2009)
33. Pan, S., Yang, Q.: A survey on transfer learning. IEEE Transactions on Knowledge and Data Engineering (TKDE) **22**(10) (2010)
34. Pardoe, D., Stone, P.: Boosting for regression transfer. In: International Conference on Machine Learning (ICML) (2010)
35. Patterson, D.: President's words used to create "deepfakes" at davos. Video (2020)
36. Peng, X., Bai, Q., Xia, X., Huang, Z., Saenko, K., Wang, B.: Moment matching for multi-source domain adaptation. In: IEEE International Conference on Computer Vision (ICCV) (2019)
37. Pumarola, A., Agudo, A., Martinez, A., Sanfeliu, A., Moreno-Noguer, F.: Ganimation: Anatomically-aware facial animation from a single image. In: European Conference on Computer Vision (ECCV) (2018)
38. Rossler, A., Cozzolino, D., Verdoliva, L., Riess, C., Thies, J., Niessner, M.: Faceforensics: A large-scale video dataset for forgery detection in human faces. In: arXiv:1803.09179 (2018)
39. Rossler, A., Cozzolino, D., Verdoliva, L., Riess, C., Thies, J., Niessner, M.: Faceforensics++: Learning to detect manipulated facial images. In: IEEE/CVF International conference on computer vision (ICCV) (2019)
40. Salloum, R., Ren, Y., Kuo, C.J.: Image splicing localization using a multi-task fully convolutional network (MFCN). J. Vis. Commun. Image Represent. **51**, 201–209 (2018)
41. Seraj, S., Singh, A., Chakraborty, S.: Semi-supervised deep domain adaptation for deepfake detection. In: Workshop on Manipulation, Adversarial, and Presentation Attacks in Biometrics at IEEE Winter Conference on Applications of Computer Vision (WACV-W) (2024)
42. Shen, J., Qu, Y., Zhang, W., Yu, Y.: Wasserstein distance guided representation learning for domain adaptation. In: Association for the Advancement of Artificial Intelligence (AAAI) (2018)
43. Shui, C., Chen, Q., Wen, J., Zhou, F., Gagne, C., Wang, B.: A novel domain adaptation theory with jensen-shannon divergence. Knowledge-Based Systems **257** (2022)
44. Sun, B., Feng, J., Saenko, K.: Return of frustratingly easy domain adaptation. In: AAAI Conference on Artificial Intelligence (2016)
45. Suwajanakorn, S., Seitz, S., Kemelmacher-Shlizerman, I.: Synthesizing obama: learning lip sync from audio. ACM Transactions on Graphics **36**(4) (2017)
46. Tariq, S., Lee, S., Woo, S.S.: One detector to rule them all: Towards a general deepfake attack detection framework. In: Proceedings of the web conference (WWW). pp. 3625–3637 (2021)
47. Thies, J., Zollhofer, M., Stamminger, M., Theobalt, C., Niessner, M.: Face2face: Real-time face capture and reenactment of RGB videos. In: IEEE conference on computer vision and pattern recognition (CVPR) (2016)
48. Thompson, B., Gwinnup, J., Khayrallah, H., Duh, K., Koehn, P.: Overcoming catastrophic forgetting during domain adaptation of neural machine translation. In: North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). pp. 2062–2068 (2019)
49. Tzeng, E., Hoffman, J., Saenko, K., Darrell, T.: Adversarial discriminative domain adaptation. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2017)

50. Venkateswara, H., Eusebio, J., Chakraborty, S., Panchanathan, S.: Deep hashing network for unsupervised domain adaptation. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2017)
51. Vincent, J.: Watch jordan peele use ai to make barack obama deliver a psa about fake news. The Verge **17** (2018)
52. Xu, R., Chen, Z., Zuo, W., Yan, J., Lin, L.: Deep cocktail network: Multi-source unsupervised domain adaptation with category shift. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2018)
53. Xu, Y., Zhong, X., Yepes, A., Lau, J.: Forget me not: Reducing catastrophic forgetting for domain adaptation in reading comprehension. In: IEEE International Joint Conference on Neural Networks (IJCNN). pp. 1–8. IEEE (2020)
54. Yang, X., Li, Y., Lyu, S.: Exposing deep fakes using inconsistent head poses. In: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 8261–8265 (2019)
55. Zellinger, W., Grubinger, T., Lughofer, E., Natschlager, T., Saminger-Platz, S.: Central moment discrepancy (cmd) for domain-invariant representation learning. In: International Conference on Learning Representations (ICLR) (2017)
56. Zhang, M., Wang, H., He, P., Malik, A., Liu, H.: Improving GAN-generated image detection generalization using unsupervised domain adaptation. In: IEEE International Conference on Multimedia and Expo (ICME) (2022)
57. Zhao, H., Zhang, S., Wu, G., Moura, J., Costeira, J., Gordon, G.: Adversarial multiple source domain adaptation. In: Advances of Neural Information Processing Systems (NeurIPS) (2018)
58. Zhong, W., Tang, D., Xu, Z., Wang, R., Duan, N., Zhou, M., Wang, J., Yin, J.: Neural deepfake detection with factual structure of text. In: Empirical Methods in Natural Language Processing (EMNLP) (2020)
59. Zhou, P., Han, X., Morariu, V.I., Davis, L.S.: Two-stream neural networks for tampered face detection. In: IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). pp. 1831–1839. IEEE (2017)
60. Zhou, P., Han, X., Morariu, V.I., Davis, L.S.: Learning rich features for image manipulation detection. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 1053–1061 (2018)

SPI2I: Structure-Preserved Image-to-Image Translation with Diffusion Models

Beibei Dong, Bo Peng, and Jing Dong(✉)

New Laboratory of Pattern Recognition (NLPR), Institute of Automation,
Chinese Academy of Sciences, Beijing 100190, China
dongbeibei2022@ia.ac.cn, {bo.peng,jdong}@nlpr.ia.ac.cn

Abstract. Large-scale text-to-image generative models are already proficient at producing high-quality results that closely match the intended prompts. Nevertheless, the pivotal challenge in image editing tasks lies in the difficulty of confining alterations within the editing region while preserving the structure and details of the source image. In this paper, we propose a zero-shot structure-preserved image-to-image translation approach based on diffusion models. We combine the optimization of the latent code and the injection of the U-Net features to strengthen the structural preservation effect by alleviating the inconsistency between the information contained in the latent code and the injected features. Our method effectively preserves the structural and detailed information of the source image while enhancing the quality of the generated results. We exhibit comprehensive and high-quality experimental results showcasing that our approach surpasses state-of-the-art methods across various image-to-image translation tasks.

Keywords: Image-to-image translation · Image editing · Text-to-image generation

1 Introduction

Turning mountains to seas, transforming photos into colorful oil pastels, or changing the autumn landscape to a snowy wonderland... as long as you can imagine it, everything can be made true with generative models. Image-to-image translation holds an immense perspective in the industrial production and design fields. GAN-based methods [10,13,20,35,39] have made remarkable strides in the field of image-to-image translation. These methods always require training specific models for an individual task, and the generated quality and resolution are limited. Owing to the powerful generation capability and impressive generation quality, diffusion models have found widespread applications in a variety of image generation tasks. Some methods [4,16,30,32,36] have explored training diffusion models for unpaired image translation. Text-guided diffusion models [19,26,29] allow introducing text conditions to guide the generation process, which enables

A. Antonacopoulos et al. (Eds.): ICPR 2024, LNCS 15321, pp. 143–159, 2025.
https://doi.org/10.1007/978-3-031-78305-0_10

more flexible image editing. Utilizing the pre-trained diffusion models, it is possible to generate high-quality results without any model training or fine-tuning. However, simply using DDIM inversion [31] with text guidance is insufficient to preserve the structure and details of the source image, which is crucial for image-to-image translation tasks. Recent studies have enhanced the structural preservation by optimizing the latent code [17,21] or replacing the intermediate features in U-Net [7,34]. Although these methods can maintain the overall structure of the source image to some extent, the details and fine structures like the texture and poses are always missed, as shown in Fig. 1.

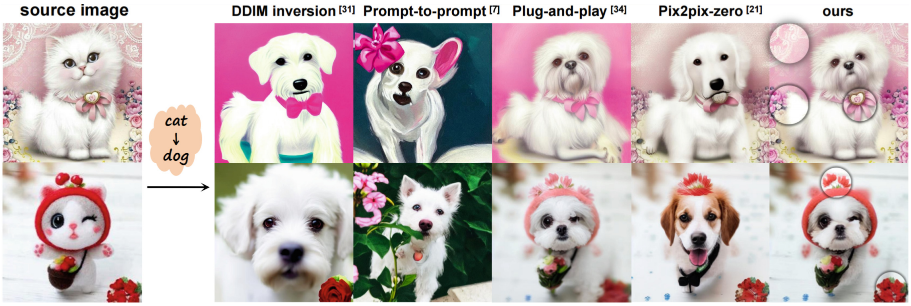


Fig. 1. Results of existing methods and our *SPI2I*. Existing diffusion-based methods struggle to transfer the structure and background details of the source image to the editing results. Our method can not only preserve the structure but also transfer the details like the texture of the source image, as circled in our results.

To enhance structural preservation, the following aspects should be considered: First, providing accurate textual descriptions is necessary. Recent methods often utilize text-to-image generation models, where text conditions not only introduce the translation objective but also provide content information for the generated images. Some methods [4,12,36] only use a pair of words like ”cat” and ”dog” to describe the source and target domains, which overlooks the crucial structural and detailed information introduced through the guiding text, as shown in Fig. 2 (a). Second, although injecting self-attention features [34] has been proven to preserve the original image structure while maintaining editability, it falls short in transferring detailed structures and information like colors and layouts. This is caused by the inconsistency between the information contained in the injected features and the noised latent code. In contrast, the method optimizing the latent code [21] shows better performance in detail transfer and perception quality but lacks the preservation of the overall structure, as shown in Fig. 2 (b).

In this paper, we introduce a novel diffusion-based framework called *SPI2I* to accomplish structure-preserved image-to-image translation without any model training or extra user-provided conditions. Given a source image and a specific editing task, our method only requires one reconstruction process to extract

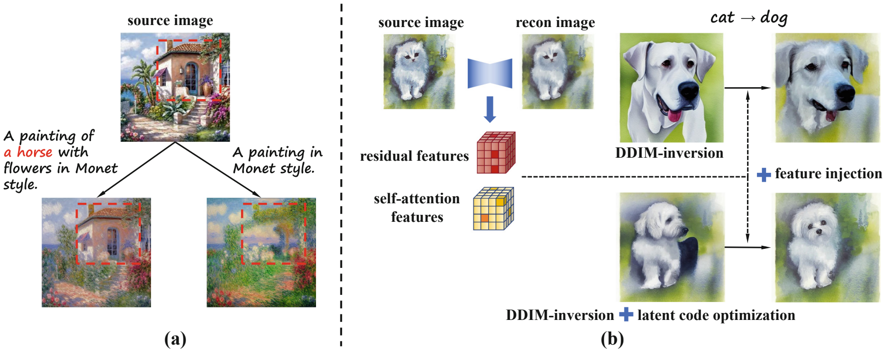


Fig. 2. Motivations. As shown in (a), when translating the source image into Monet style, if the information "house" is not provided in the text, the generated result will miss this important content. In Figure (b), when injecting the features of the reconstruction process into the reverse process of DDIM inversion, the inconsistency between the structural information contained in the injected features and the latent code of DDIM inversion leads to a decline in generation quality. In contrast, optimizing the latent code before injecting features can both enhance generation quality and improve structural preservation.

reference features. Then at each timestep of the editing process, we first optimize the latent code reliant on the reference cross-attention features to incorporate a certain degree of structure and details from the source image. After the optimization, we selectively inject the reference features of the residual and self-attention blocks into the editing process. We also employ BLIP [14] as an automatic description extractor to provide accurate prompts about the content of the source images.

Our main contributions are: 1) We propose *SPI2I*, a diffusion-based framework for zero-shot image-to-image translation, to enhance the structural preservation effect using optimization-based and feature injection methods. 2) By optimizing the latent code of diffusion, we alleviate the inconsistency between the information contained in the injected features and the latent code, thus improving the perception quality. 3) We employ BLIP [14] to extract precise editing prompts, which enhances editing accuracy while allowing large-scale automatic editing. The experiment results show that our method achieves an outstanding structural preservation effect, producing high-quality results that align well with the target prompt, outperforming state-of-the-art methods in multiple metrics.

2 Related Work

Text-to-image generative models. Due to the versatile capabilities of the multimodal model CLIP [23] aligning the image and text latent spaces, recent image generative models like GANs and diffusion models tend to exploit the CLIP model to enhance editing control. For example, Ramesh et al. [24,25]

utilized a transformer as a prior model, aligning text and image latent spaces using CLIP, and presented a CLIP ranking technique. Recent investigations have focused on leveraging diffusion models to generate high-quality results. Nichol et al. [19] investigated a Noised-CLIP-guided conditional generation method and proposed classifier-free guidance, allowing for diverse conditional guidance without additional classifiers. Latent diffusion models [26] reverse the image into the latent space for the sampling process, saving the computational overhead while enabling text-guided generation with CLIP. Recent explorations aim to introduce enhanced control over the generation process using various forms of conditions by training hypernetworks [18,37] or optimizing the latent code [28, 28], expanding the application scope of text-to-image generation models.

Image editing with diffusion models. Existing methods [19,24,29] can generate high-quality results and guide the generation by introducing text conditions. However, image editing tasks require stricter structural control. Kawar et al. [11] proposed a fine-tuning method to accomplish a specific editing task on a single image. Instruct-Pix2pix [4] utilizes GPT-3 to train the diffusion models to edit images based on the given instruction. To avoid the additional training overhead, Some work [1,2,15] offer extra conditions such as masks to enhance controllability in image editing tasks. Other methods [6,7,34] extract and visualize the intermediate features of diffusion models, analyzing the effects of different features on structural preservation, thus transferring the structural information by directly manipulating the features.

Image-to-image translation. Image-to-image translation aims to edit an image from the source domain to the target domain while retaining other structures and details. Earlier GAN-based methods [20,35,39] train on paired images by reconstructing the input image. To alleviate the dependency on paired data, unsupervised image-to-image translation methods [10,13] adopt cycle consistency loss to achieve domain translation. Recently, some methods [4,16,30,32,36] have explored training diffusion models for unpaired image translation. However, these approaches demand considerable computational resources and time overhead on a specific task. DiffuseIT [12] incorporates CLIP Similarity and DINO-ViT Similarity-based loss to disentangle the content and domain attributes and optimize the latent code. Parmar et al. [21] leveraged GPT-3.5 to discover the edit direction in the embedding space and optimized the latent code based on the cross-attention maps. Although existing methods achieve zero-shot image-to-image translation on multiple tasks, their capability of producing high-quality images while preserving the original structure is still insufficient.

3 Preliminary

Diffusion models are probabilistic generative models that estimate noise in the input data to recover the data distribution from the Gaussian noise. During the DDIM forward process, the model adds random noise into an initial image, yielding a noisy image $\boldsymbol{x}_t$:

$$\boldsymbol{x}_t = \sqrt{\bar{\alpha}_t} \cdot \boldsymbol{x}_0 + \sqrt{1 - \bar{\alpha}_t} \cdot \boldsymbol{z} \tag{1}$$

where $\boldsymbol{z} \sim \mathcal{N}(0, \mathbf{I})$ and $\{\bar{\alpha}_t\}$ are noise scheduler.

In the backward process, the model estimates the noise introduced at each timestep of the forward process, gradually denoising and recovering the input image.

$$\boldsymbol{x}_{t-1} = \frac{\boldsymbol{x}_t - \sqrt{1-\alpha}_t \cdot \boldsymbol{\epsilon}_\theta(\boldsymbol{x}_t, c, t)}{\sqrt{\alpha_t}} \tag{2}$$

$\alpha_t = \frac{\bar{\alpha}_t}{\bar{\alpha}_{t-1}}$, $\boldsymbol{\epsilon}_\theta(\boldsymbol{x}_t, t, c)$ is the estimated noise at timestep t. c is the condition used to control the generation process.

Classifier-free guidance [9] introduces conditional guidance into the generative process without additional classifiers. Given the condition c and an empty condition $\varnothing$, classifier-free guidance is represented as follows:

$$\boldsymbol{\epsilon} = \boldsymbol{\epsilon}_\theta(\boldsymbol{x}_t, \varnothing, t) + s \cdot (\boldsymbol{\epsilon}_\theta(\boldsymbol{x}_t, c, t) - \boldsymbol{\epsilon}_\theta(\boldsymbol{x}_t, \varnothing, t)) \tag{3}$$

where $s \in [0, 1]$ is the scaling factor. We use $\boldsymbol{\epsilon}$ instead of $\boldsymbol{\epsilon}_\theta(\boldsymbol{x}_t, c, t)$ in Eq. (2) in classifier-free guidance during the generation process.

Diffusion models contain a U-Net architecture [27]. Residual layers, self-attention layers, and cross-attention layers are arranged in sequence in U-Net. Given query matrix Q, key matrix K, and value matrix V, the output of the attention layer is represented by:

$$\text{Attention}(Q, K, V) = M \cdot V, \text{ where } M = \text{Softmax}\left(\frac{QK^T}{\sqrt{d}}\right) \tag{4}$$

The matrices Q, K, and V are obtained by applying the corresponding learned projections on the input features. d is the dimension of the projection keys and values. Features can refer to either spatial features or text embeddings according to the type of attention layer.

4 Method

Our method first extracts accurate text descriptions from the input image. Then, we employ two different structural preservation methods, named refined cross-attention guidance and feature injection, to generate high-quality editing results with exceptional structural preservation effects.

4.1 Prompt Extraction

Utilizing text descriptions to guide the generation process provides control over the content of the image generated by diffusion models, therefore describing the editing target accurately is essential for achieving precise editing.

In latent diffusion models, text conditions are first encoded into the CLIP text embedding space, and then introduced into the cross-attention blocks in the U-Net architecture. Using the image caption model BLIP [14], we extract text descriptions as the representation of the source image content. Moreover, using

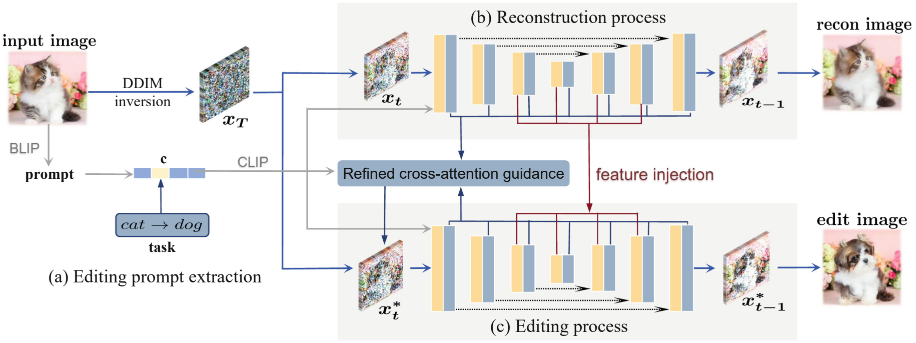


Fig. 3. Pipeline of our *SPI2I*. Given an input image and an editing task, such as *cat*→*dog*, we guide the generation process with multiple structural constraints. Our pipeline primarily contains three parts. (a) Extracting prompt using BLIP model. (b) Extracting reference features during the reconstruction process. (c) Directing the editing process using structural constraints. Our method mainly leverages two different structural preservation methods, named feature injection and refined cross-attention guidance.

the image caption model instead of the manual annotation allows fully automatic editing of a large number of images, which is useful in practical scenarios. After extracting the description of the source image using BLIP, we swap or add the target words in terms of the specified task into the text as the prompt guiding the editing process.

4.2 Refined Cross-attention Guidance

The U-Net in latent diffusion models comprises cross-attention blocks at each module. Text condition is introduced through these cross-attention blocks. At timestep t, the output of the cross-attention layer is given by Eq. (4), where matrix Q is acquired from the spatial features, and matrices K, V are derived from text embeddings. Cross-attention guidance manipulates the cross-attention blocks using an optimization-based method. By iteratively optimizing the initial latent at each timestep using a loss named $\mathcal{L}_{\text{cag}}$, this method can effectively retain the structural and detailed information of the source image [21]. Our method is inspired by this method.

Initially, we employ the description extracted from the reference image as text conditions for the reconstruction of the reference image. During the reconstruction process, a series of cross-attention maps from different indices of blocks, denoted as $\left\{\boldsymbol{M}_t^{ref}\right\}$ at timestep t, are obtained (see Fig. 4 (b)). The last dimension of each $\boldsymbol{M}_t^{ref}$ corresponds to the length of the text condition tokens, with each channel of this dimension representing the cross-attention feature of the corresponding token.

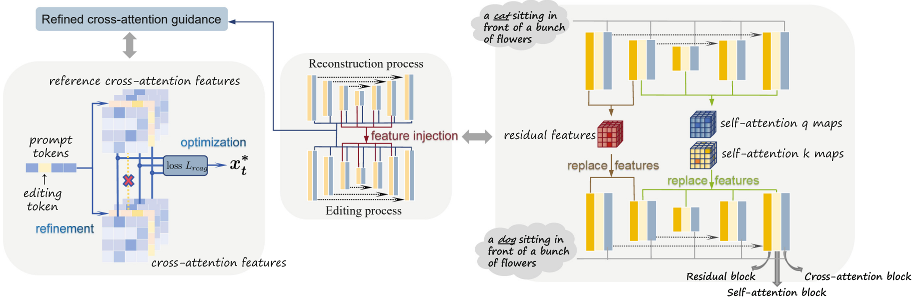


Fig. 4. Refined cross-attention guidance and feature injection. In refined cross-attention guidance, we utilize reference cross-attention features extracted during the reconstruction process to optimize the latent space representation at each timestep, the optimized representation at timestep t is denoted by x_t^*. Afterward, we directly replace the selected residual block features and self-attention maps of the editing process with the reference features of the reconstruction process, named feature injection.

Then, in the editing process, we obtain a series of cross-attention maps, denoted as $\left\{\boldsymbol{M}_t^{edit}\right\}$. Feature $\boldsymbol{M}_t^{edit}$ are then used to compute the cross-attention guidance loss $\mathcal{L}_{\text{cag}}$ in relation to the corresponding $\boldsymbol{M}_t^{ref}$ derived from the reconstruction process, as expressed by:

$$\mathcal{L}_{\text{cag}} = \sum_{index} \left\| \boldsymbol{M}_t^{\text{edit}} - \boldsymbol{M}_t^{\text{ref}} \right\|_2 \tag{5}$$

$\sum_{index}$ represents the sum of losses in different indices of cross-attention blocks.

Finally, loss $\mathcal{L}_{\text{cag}}$ is used to optimize the initial latent code at timestep t. However, this structural constraint relatively results in decreased editability within the editing region, consequently diminishing the accuracy of the generated results. We introduce a novel refinement method within the cross-attention guidance to alleviate this decline. As depicted in Fig. 4 refined cross-attention guidance, considering a specific task, we can locate the editing tokens within the tokens of the prompt, and then identify the corresponding channels within the cross-attention map associated with the edited terms in the prompt. When calculating optimization loss $\mathcal{L}_{\text{cag}}$, we exclude these maps, thereby deviating from the source image in the editing region, while preserving the background structure. We use a cross-attention mask to represent this refinement. The refined cross-attention guidance loss $\mathcal{L}_{\text{rcag}}$ is given by:

$$\mathcal{L}_{\text{rcag}} = \sum_{index} \left\| \boldsymbol{m} \odot \left(\boldsymbol{M}_t^{\text{edit}} - \boldsymbol{M}_t^{\text{ref}} \right) \right\|_2 \tag{6}$$

$\boldsymbol{m}$ is the cross-attention mask obtained from the text tokens, where the positions corresponding to tokens changed for transferring from the source to the target domain are set to 0, and other positions are set to 1. $\odot$ donates the channel-wise product between the mask $\boldsymbol{m}$ and cross-attention maps.

4.3 Feature Injection

Semantic layout encapsulates essential structural information of the corresponding images. Many studies [1,2,22] have sought to achieve precise control over image editing regions by introducing various forms of semantic representations. Baranchuk et al. [3] have proved that the spatial features of unconditional diffusion models contain crucial semantic information, and thus exhibit a close connection with the structure of the generated results. In [7,34], PCA analysis is used to visualize spatial features from different layers of the diffusion model at different timesteps, thereby confirming the relationship between the generated image attributes and the features.

At timestep t in the editing process, we override the residual block features with the corresponding features $\boldsymbol{f}_t^l$ derived from the reconstruction process, we add a * to distinguish the modified estimated noise from the original one at timestep t:

$$\boldsymbol{z}_{t-1}^* = \boldsymbol{\epsilon}_{\boldsymbol{\theta}}^* \left(\boldsymbol{x}_t^*, c, t; \left\{ \boldsymbol{f}_t^l \right\} \right) \tag{7}$$

$\boldsymbol{\epsilon}_{\boldsymbol{\theta}}^* \left(\cdot; \left\{ \boldsymbol{f}_t^l \right\} \right)$ denotes the modified noise output of the denoising process with the injected features $\left\{ \boldsymbol{f}_t^l \right\}$ from residual block indices $\{l\}$.

Injecting self-attention maps into the editing process is proved to improve the affinities between the spatial features and target semantics [34]. In residual blocks, we directly replace the specific features with the injected features, while the injection operation in self-attention blocks differs slightly. The self-attention layer is defined in Eq. (4), and all matrices are computed on the input spatial features. In our method, during the editing process, the query matrix Q and key matrix K are substituted with the matrices from the reconstruction process. Incorporating with the self-attention injection, the output of the U-Net can be represented as follows:

$$\boldsymbol{z}_{t-1}^* = \boldsymbol{\epsilon}_{\boldsymbol{\theta}}^* \left(\boldsymbol{x}_t^*, c, t; \left\{ \boldsymbol{f}_t^l \right\}; \left\{ \boldsymbol{A}_t^h \right\} \right) \tag{8}$$

$\left\{ \boldsymbol{A}_t^h \right\}$ is the injected maps, corresponding to matrices Q, K in Eq. (4), obtained from the reconstruction process at timestep t. h represents the self-attention block index.

Feature injection directly incorporates structural information from the reference image into the editing process, thereby ensuring robust structural preservation. However, using only the feature injection method can lead to decreased generation quality, and the color and texture details transferring is unsatisfactory. In practice, we first utilize the refined cross-attention guidance to optimize the initial latent code at each timestep, then proceed with the feature injection method during the editing process. This arrangement enhances the affinity between the injected features and the original features of the editing process, thereby alleviating the decreased generation quality caused by feature injection, and the feature injection method also helps avoid the blurring introduced by our optimization-based method in Sec. 4.2. The algorithm and implementation details are provided in Appendix A.

5 Results

We apply our method described in Sec. 4 to diverse tasks and compare it with the state-of-the-art zero-shot image-to-image translation methods. For each task to translate the input image into the target domain, we select 200 images that belong to the source domain from LAION-5B dataset. We apply our method on real and synthetic images, following the setup in our baseline Pix2pix-zero [21]. For the synthetic image dataset, we employ Stable Diffusion to generate 200 images that belong to the source domain according to the prompts generated by GPT-3.5, then translate these synthetic images to the target domain. Some results are shown in Fig. 5. Implementation details and more results are shown in Appendix.

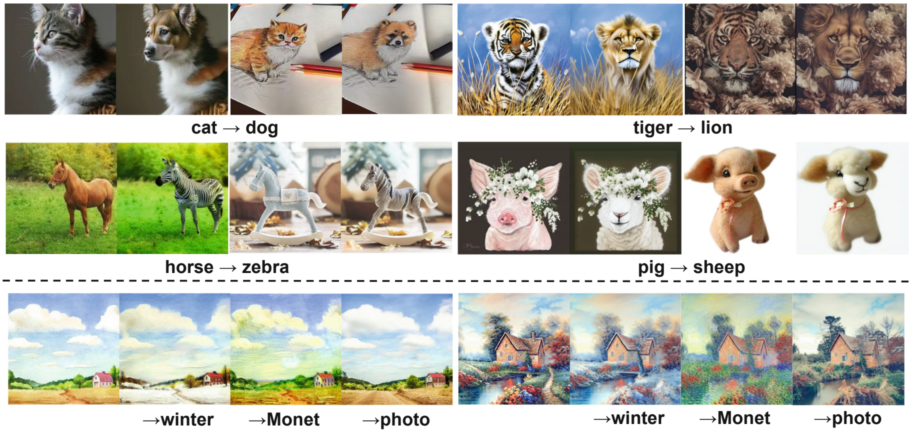


Fig. 5. Results of our *SPI2I*. We apply our method to diverse cross-domain image-to-image translation tasks. Given a source image and a specific editing task, our method can generate high-quality results aligned with the editing target while preserving the structure of the source images without model training or fine-tuning.

5.1 Evaluation and Comparison

Metrics. We conduct quantitative evaluations of our method and the comparative methods using multiple metrics, covering editing accuracy and structural preservation. For editing accuracy, we utilize CLIP cosine similarity [8] to measure the similarity between the generated image and the editing prompt, where a higher value indicates a better resemblance between the generated image and the prompt. To evaluate the effect of structural preservation, two metrics are employed: the perceptual similarity [38] and the DINO self-similarity [33]. For perceptual similarity, we use the AlexNet-based LPIPS to assess the structural similarity between the generated image and the source image. A smaller LPIPS value indicates a closer structural resemblance between the two images. DINO self-similarity employs the DINO-ViT model to extract image patch features and compute similarity, a smaller value indicates better structural preservation. To further evaluate the quality of generated images, we conduct a user study and present the result in Appendix C.

We compare our method against state-of-the-art zero-shot image-to-image translation methods, including DiffuseIT [12], Prompt-to-prompt [7], Plug-and-play [34], MasaCtrl [6] and Pix2pix-zero [21]. All experiments do not require additional model training or fine-tuning. For a fair comparison, we employ the same sampling steps and precision in all experiments. As a specific note, Pix2pix-zero utilizes a technique involving GPT-3 [5] to derive the editing direction within the CLIP text embedding space. Because this method is also applicable in our approach, and we lack the explicit details regarding this method, we slightly

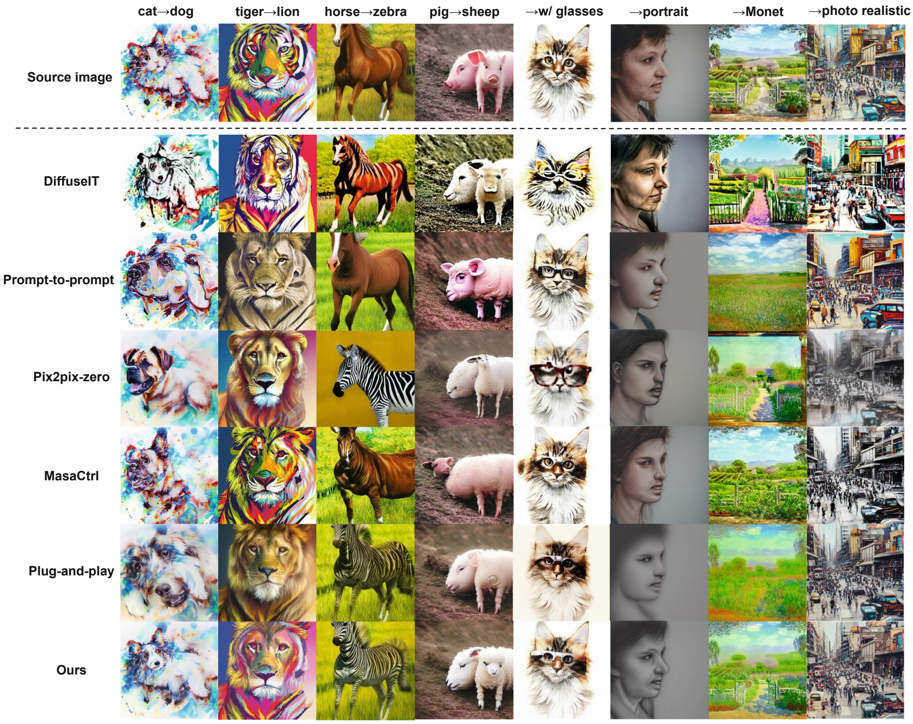


Fig. 6. Comparison with state-of-the-art methods. We compare our method against state-of-the-art approaches across various tasks. Our method delivers high-quality results, ensuring accurate adherence to the editing target and strong structural alignment with the source image, surpassing other approaches.

modify the method finding editing direction in Pix2pix-zero to word swap, which is the same as ours. For tasks editing the whole image, like transferring an image to Monet style, we insert the editing target words, such as Monet style, to the prompt.

Qualitative results are presented in Fig. 6. As illustrated, our method achieves high-quality editing outputs and demonstrates fidelity to the editing objectives, outperforming state-of-the-art methods. Tab. 1 displays the quantitative evaluations, comparing our approach with state-of-the-art methods by (i) CLIP-Sim(CLIP cosine similarity), (ii) LPIPS(Perceptual similarity) and (iii) DINO-Sim(DINO self-similarity).

Table 1. Evaluation and Comparison. We evaluate our method on six different tasks, including three real image editing tasks in sub-table (a) and synthetic image editing tasks in sub-table (b). Our method surpasses all other methods in structural preservation with the best LPIPS scores, showcasing superior structural preservation with satisfactory editing accuracy. As a specific note, we dismiss the DINO-Sim score of DiffuseIT as it employs this metric within its optimization loss.

	cat→dog			cat→cat with glasses			Monet		
	CLIP-Sim↑	LPIPS↓	DINO-Sim↓	CLIP-Sim↑	LPIPS↓	DINO-Sim↓	CLIP-Sim↑	LPIPS↓	DINO-Sim↓
DiffuseIT	0.6760	0.4036	**0.0458**	0.7521	0.3989	**0.0462**	0.6765	0.4523	**0.0551**
Prompt-to-prompt	0.6206	0.3740	0.0804	0.7382	0.3428	0.0786	0.7612	0.3913	0.0744
Pix2pix-zero	0.7628	0.3915	0.0813	0.8314	0.3607	0.0776	0.7228	0.3685	0.0749
MasaCtrl	0.6278	0.3407	0.0859	0.7149	0.3327	0.0851	0.6823	0.3521	0.0800
Plug-and-play	0.6734	0.3968	0.0826	0.7788	0.3768	0.0797	0.7623	0.4965	0.0866
ours	**0.7780**	**0.3401**	0.0741	**0.8369**	**0.3187**	0.0737	**0.7425**	**0.3263**	0.0684

(a) Real image editing

	pig→sheep			painting→photo			photo→portrait drawing		
	CLIP-Sim↑	LPIPS↓	DINO-Sim↓	CLIP-Sim↑	LPIPS↓	DINO-Sim↓	CLIP-Sim↑	LPIPS↓	DINO-Sim↓
DiffuseIT	0.6903	0.3884	0.0376	0.5749	0.4144	0.0483	0.5487	0.3663	0.0456
Prompt-to-prompt	0.7512	0.2420	0.0408	0.7763	0.1979	0.0379	0.7984	0.1803	0.0344
Pix2pix-zero	0.7955	0.2879	0.0454	0.7803	0.2431	0.0408	0.8080	0.2286	0.0429
MasaCtrl	0.7936	0.2411	0.0468	0.7817	0.1888	0.0357	0.8058	0.1932	0.0407
Plug-and-play	0.7898	0.2420	0.0320	0.7950	0.2020	0.0386	**0.8165**	0.2190	0.0304
ours	**0.7956**	**0.2382**	**0.0276**	**0.7955**	**0.1867**	**0.0348**	0.8081	**0.1599**	**0.0207**

(b) Synthetic image editing

As shown in Tab. 1, our method achieves the best structural preservation indicated by the lowest LPIPS scores. An essential clarification is that, due to the utilization of DINO self-similarity as the optimization loss in DiffuseIT, the DINO-Sim score of DiffuseIT is naturally lower than other methods. As a result, we think the DINO-Sim score of the DiffuseIT method does not have much significance. Our method also shows fidelity to the target word, higher than other methods in terms of the CLIP-Sim metric.

5.2 Ablation Study

Our approach primarily includes three techniques for structural preservation, including refined cross-attention guidance, feature injection and extracting precise prompt using BLIP [14], as proposed in Sec. 4. To further substantiate

Table 2. Quantitative evaluation of the ablation study. We combine different components of our method separately in each group from config A to D. Our full method in config E achieves the best structural preservation by the lowest LPIPS.

	cat→dog		tiger→lion		cat→cat w/ glasses		Monet	
	CLIP-Sim ↑	LPIPS ↓	CLIP-Sim ↑	LPIPS ↓	CLIP-Sim ↑	LPIPS ↓	CLIP-Sim ↑	LPIPS ↓
A w/ cag	0.7628	0.3915	0.7620	0.4296	0.8314	0.3607	0.7228	0.3685
B w/ refined cag	**0.8109**	0.4500	**0.7972**	0.4806	**0.8878**	0.4176	**0.8153**	0.4472
C w/ inj.	0.8024	0.3870	0.7934	0.4498	0.8551	0.3624	0.8105	0.4376
D w/o prompt extraction	0.6406	0.3616	0.6027	0.4070	0.7462	0.3450	0.7624	0.4821
E w/ refined cag and inj.(**ours**)	0.7818	**0.3428**	0.7749	**0.3844**	0.8369	**0.3187**	0.7691	**0.3607**

our method, we conduct an ablation study where we combine different components of our method separately in each group, including (A)with cross-attention guidance, (B)with refined cross-attention guidance, (C)with feature injection, (D)without prompt extraction, and finally our method with all components in config E. We evaluate the consistency between the generated images and the target prompt using CLIP-Sim, evaluating the structural preservation effect using LPIPS scores in four different tasks.

Quantitative evaluation is presented in Tab. 2. In config A, we use the vanilla cross-attention guidance, and then our refinement method is applied in config B. We can see significant improvement in CLIP-Sim scores, indicating enhanced editability. Config C shows the result of only using the feature injection method. In config D, we simply use a prompt like ”a photo of a cat” instead of extracting precise prompts using BLIP [14]. Our full method in config E achieves the best structural preservation effect by the lowest LPIPS.

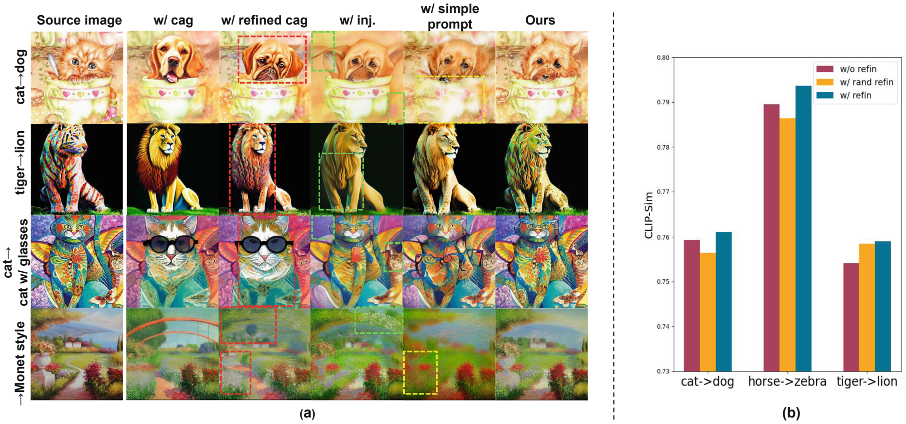
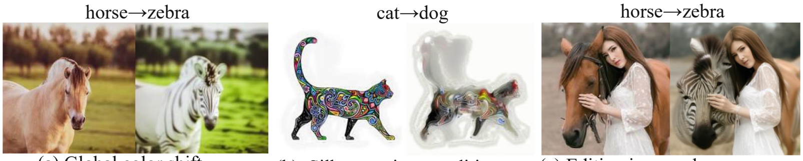


Fig. 7. Results of the ablation study. As shown in sub-figure (a), removing the feature injection method (indicated by *w/ refined cag.*) results in decreased preservation of overall structure and posture, while only using it incurs significant losses in the preservation of details like color and texture, as highlighted by the green boxes in *w/ inj.* We further conduct an experiment to demonstrate the enhanced editability brought by our refinement method in cross-attention guidance, results shown in sub-figure (b).

Direct removal of feature injction(config B) and refined cross-attention guidance(config C) from our full method(config E) can lead to a noticeable decline in structural preservation. As shown in Fig. 7 (a), while the feature injection method has a strong effect on structure preservation, it shows significant losses in background details such as texture, which can be addressed by the refined cross-attention guidance method. In contrast, using only the refined cross-attention guidance method results in poor effectiveness in structure preservation. However, the results of config B and C show better CLIP-Sim metrics than our final setting. Image-to-image translation involves a trade-off between editability and fidelity. In our method, structural constraints relatively weaken the flexibility of the editing process guided by text condition, thereby leading to a decrease in the CLIP-Sim score. Our method does not exhibit a substantial decline in CLIP-Sim score while preserving the overall structure and details of the source image, as shown in Fig. 7 (a). We further conduct experiments to investigate the trade-off between editability and fidelity in Appendix B.

Comparing config A and B on all tasks, our refinement method in cross-attention guidance demonstrates a significant improvement in the CLIP-Sim metric. To verify that this improvement is not merely a result of reducing structural constraints, we compare our method, donated as *w/ refin*, with randomly masking the same amount of channels in the cross-attention maps with our refinement method, denoted as *w/ rand refin*. Our method with vanilla cross-attention guidance is denoted as *w/o refin*. The results are shown in Fig. 7 (b). The highest CLIP-Sim scores across all tasks indicate that our refinement method in cross-attention guidance can enhance the accuracy between the generated result and the editing objective. The results of the user study in Appendix C can also prove this.

In config D, we use simple prompts instead of using BLIP [14] to extract precise text descriptions of the source image. Both the CLIP-Sim and LPIPS scores worsen compared with our full method, showcasing the importance of the prompt extraction method in enhancing text alignment and structural preservation. Some results of config D are shown in Fig. 7 *w/ simple prompt*. We can observe the loss of some background information due to the imprecise text descriptions, as highlighted by the yellow boxes.



(a) Global color shift. (b) Silhouette image editing. (c) Editing in complex scenes.

Fig. 8. Limitations. We show some cases where our method generates unsatisfying results and discuss.

6 Discussion and Conclusion

We introduce a novel image-to-image translation framework called *SPI2I*. Given a single input image along with a specified editing task, our method can generate high-quality editing results, retaining the structure and background details of the source image without model training or fine-tuning. Motivated by the inconsistency between the information contained in the latent code and the injected features, our approach involves refined cross-attention guidance and a feature injection method, ensuring robust structural preservation while accurately aligning with the editing target. We conduct adequate experiments and substantiate that our method achieves strong structural preservation and high-quality generation. We also observed some shortcomings in our experiments.

As shown in Fig. 8 (a), We observed global color shifts in some results. We further substantiate that this outcome is associated with the feature injection method by comparing it with the result without feature injection. Optimization of the latent code can alleviate it to some extent.

Fig. 8 (b) and (c) illustrate some bad results. For silhouette image editing, the semantic layout is hard to locate, thus generating unsuccessful results. Fig. 8 (c) displays scenes involving interactions between editing target objects and other elements. Editing in complex scenes remains a persistent challenge. Introducing additional constraints, such as mask conditions, can help address these issues, but also leads to limited applicability at the same time.

Overall, our method is zero-shot and the results show that it can produce high-quality images, tackling the issue of structural preservation in image-to-image translation and image editing, outperforming state-of-the-art methods across multiple metrics.

Acknowledge. This work is supported by the National Key Research and Development Program of China under Grant No. 2021YFC3320103, the National Natural Science Foundation of China (NSFC) under Grant No. 62272460, Beijing Natural Science Foundation under Grant No. 4232037.

References

1. Avrahami, O., Fried, O., Lischinski, D.: Blended latent diffusion. ACM Transactions on Graphics (TOG) **42**(4), 1–11 (2023)
2. Avrahami, O., Lischinski, D., Fried, O.: Blended diffusion for text-driven editing of natural images. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 18208–18218 (2022)
3. Baranchuk, D., Rubachev, I., Voynov, A., Khrulkov, V., Babenko, A.: Label-efficient semantic segmentation with diffusion models. arXiv preprint arXiv:2112.03126 (2021)
4. Brooks, T., Holynski, A., Efros, A.A.: Instructpix2pix: Learning to follow image editing instructions. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 18392–18402 (2023)

5. Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J.D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al.: Language models are few-shot learners. Adv. Neural. Inf. Process. Syst. **33**, 1877–1901 (2020)
6. Cao, M., Wang, X., Qi, Z., Shan, Y., Qie, X., Zheng, Y.: Masactrl: Tuning-free mutual self-attention control for consistent image synthesis and editing. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 22560–22570 (2023)
7. Hertz, A., Mokady, R., Tenenbaum, J., Aberman, K., Pritch, Y., Cohen-Or, D.: Prompt-to-prompt image editing with cross attention control. arXiv preprint arXiv:2208.01626 (2022)
8. Hessel, J., Holtzman, A., Forbes, M., Bras, R.L., Choi, Y.: Clipscore: A reference-free evaluation metric for image captioning. arXiv preprint arXiv:2104.08718 (2021)
9. Ho, J., Salimans, T.: Classifier-free diffusion guidance. arXiv preprint arXiv:2207.12598 (2022)
10. Huang, X., Liu, M.Y., Belongie, S., Kautz, J.: Multimodal unsupervised image-to-image translation. In: Proceedings of the European conference on computer vision (ECCV). pp. 172–189 (2018)
11. Kawar, B., Zada, S., Lang, O., Tov, O., Chang, H., Dekel, T., Mosseri, I., Irani, M.: Imagic: Text-based real image editing with diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6007–6017 (2023)
12. Kwon, G., Ye, J.C.: Diffusion-based image translation using disentangled style and content representation. In: The Eleventh International Conference on Learning Representations (2022)
13. Lee, H.Y., Tseng, H.Y., Huang, J.B., Singh, M., Yang, M.H.: Diverse image-to-image translation via disentangled representations. In: Proceedings of the European conference on computer vision (ECCV). pp. 35–51 (2018)
14. Li, J., Li, D., Xiong, C., Hoi, S.: Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In: International Conference on Machine Learning. pp. 12888–12900. PMLR (2022)
15. Lugmayr, A., Danelljan, M., Romero, A., Yu, F., Timofte, R., Van Gool, L.: Repaint: Inpainting using denoising diffusion probabilistic models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11461–11471 (2022)
16. Meng, C., He, Y., Song, Y., Song, J., Wu, J., Zhu, J.Y., Ermon, S.: Sdedit: Guided image synthesis and editing with stochastic differential equations. arXiv preprint arXiv:2108.01073 (2021)
17. Mokady, R., Hertz, A., Aberman, K., Pritch, Y., Cohen-Or, D.: Null-text inversion for editing real images using guided diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6038–6047 (2023)
18. Mou, C., Wang, X., Xie, L., Zhang, J., Qi, Z., Shan, Y., Qie, X.: T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. arXiv preprint arXiv:2302.08453 (2023)
19. Nichol, A., Dhariwal, P., Ramesh, A., Shyam, P., Mishkin, P., McGrew, B., Sutskever, I., Chen, M.: Glide: Towards photorealistic image generation and editing with text-guided diffusion models. arXiv preprint arXiv:2112.10741 (2021)
20. Park, T., Liu, M.Y., Wang, T.C., Zhu, J.Y.: Semantic image synthesis with spatially-adaptive normalization. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 2337–2346 (2019)

21. Parmar, G., Kumar Singh, K., Zhang, R., Li, Y., Lu, J., Zhu, J.Y.: Zero-shot image-to-image translation. In: ACM SIGGRAPH 2023 Conference Proceedings. pp. 1–11 (2023)
22. Pernuš, M., Štruc, V., Dobrišek, S.: High resolution face editing with masked gan latent code optimization. arXiv preprint arXiv:2103.11135 (2021)
23. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PMLR (2021)
24. Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., Chen, M.: Hierarchical text-conditional image generation with clip latents. arXiv preprint arXiv:2204.06125 **1**(2), 3 (2022)
25. Ramesh, A., Pavlov, M., Goh, G., Gray, S., Voss, C., Radford, A., Chen, M., Sutskever, I.: Zero-shot text-to-image generation. In: International Conference on Machine Learning. pp. 8821–8831. PMLR (2021)
26. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10684–10695 (2022)
27. Ronneberger, O., Fischer, P., Brox, T.: U-Net: Convolutional Networks for Biomedical Image Segmentation. In: Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F. (eds.) MICCAI 2015. LNCS, vol. 9351, pp. 234–241. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-24574-4_28
28. Ruiz, N., Li, Y., Jampani, V., Pritch, Y., Rubinstein, M., Aberman, K.: Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 22500–22510 (2023)
29. Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E.L., Ghasemipour, K., Gontijo Lopes, R., Karagol Ayan, B., Salimans, T., et al.: Photorealistic text-to-image diffusion models with deep language understanding. Adv. Neural. Inf. Process. Syst. **35**, 36479–36494 (2022)
30. Sasaki, H., Willcocks, C.G., Breckon, T.P.: Unit-ddpm: Unpaired image translation with denoising diffusion probabilistic models. arXiv preprint arXiv:2104.05358 (2021)
31. Song, J., Meng, C., Ermon, S.: Denoising diffusion implicit models. arXiv preprint arXiv:2010.02502 (2020)
32. Su, X., Song, J., Meng, C., Ermon, S.: Dual diffusion implicit bridges for image-to-image translation. arXiv preprint arXiv:2203.08382 (2022)
33. Tumanyan, N., Bar-Tal, O., Bagon, S., Dekel, T.: Splicing vit features for semantic appearance transfer. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10748–10757 (2022)
34. Tumanyan, N., Geyer, M., Bagon, S., Dekel, T.: Plug-and-play diffusion features for text-driven image-to-image translation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1921–1930 (2023)
35. Wang, T.C., Liu, M.Y., Zhu, J.Y., Tao, A., Kautz, J., Catanzaro, B.: High-resolution image synthesis and semantic manipulation with conditional gans. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 8798–8807 (2018)
36. Wu, C.H., De la Torre, F.: Unifying diffusion models' latent space, with applications to cyclediffusion and guidance. arXiv preprint arXiv:2210.05559 (2022)

37. Zhang, L., Rao, A., Agrawala, M.: Adding conditional control to text-to-image diffusion models. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 3836–3847 (2023)
38. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 586–595 (2018)
39. Zhu, P., Abdal, R., Qin, Y., Wonka, P.: Sean: Image synthesis with semantic region-adaptive normalization. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 5104–5113 (2020)

FIDAVL: Fake Image Detection and Attribution Using Vision-Language Model

Mamadou Keita[1](✉), Wassim Hamidouche[2], Hessen Bougueffa Eutamene[1], Abdelmalik Taleb-Ahmed[1], and Abdenour Hadid[3]

[1] Laboratory of IEMN, Univ. Polytechnique Hauts-de-France, Valenciennes, France
mamadou.keita@uphf.fr
[2] Univ. Rennes, INSA Rennes, CNRS, IETR - UMR, 6164 Rennes, France
[3] Sorbonne Center for Artificial Intelligence, Sorbonne University Abu Dhabi, Abu Dhabi, UAE

Abstract. We introduce FIDAVL: Fake Image Detection and Attribution using a Vision-Language Model. FIDAVL is a novel and efficient multitask approach inspired by the synergies between vision and language processing. Leveraging the benefits of zero-shot learning, FIDAVL exploits the complementarity between vision and language along with soft prompt-tuning strategy to detect fake images and accurately attribute them to their originating source models. We conducted extensive experiments on a comprehensive dataset comprising synthetic images generated by various state-of-the-art models. Our results demonstrate that FIDAVL achieves an encouraging average detection accuracy of 95.42% and F1-score of 95.47% while also obtaining noteworthy performance metrics, with an average F1-score of 92.64% and ROUGE-L score of 96.50% for attributing synthetic images to their respective source generation models. The source code of this work will be publicly released at https://github.com/Mamadou-Keita/FIDAVL.

Keywords: Vision Language Model · Large Language Model · Deepfake · Image Captioning · Synthetic Image Attribution · Diffusion Models

1 Introduction

Over the past two decades, the landscape of techniques for generating and manipulating photorealistic images has undergone rapid evolution. This evolution has ushered in an era where visual content can be easily created and manipulated, leaving behind minimal perceptual traces. Consequently, there is a growing apprehension that we are on the brink of a world where distinguishing real images from computer generated ones will become increasingly challenging. Recent advancements in generative models have further propelled the quality and realism of synthesized images, enabling their application in conditional scenarios for contextual manipulation and broadening the scope of media

A. Antonacopoulos et al. (Eds.): ICPR 2024, LNCS 15321, pp. 160–176, 2025.
https://doi.org/10.1007/978-3-031-78305-0_11

synthesis. However, amidst these advancements, a prevailing concern persists regarding the potential repercussions of these technologies when wielded maliciously. This apprehension has garnered significant public attention due to its disruptive implications for visual security, legal frameworks, political landscapes, and societal norms [19]. Therefore, it is paramount to delve into the development of effective visual forensic techniques capable of mitigating the threats posed by these evolving generative patterns.

To tackle the challenges posed by generative models, several solutions have emerged in the literature. Existing methodologies predominantly revolve around binary detection strategies (real vs. AI-generated) [8,35] aimed at discerning synthetic images from authentic ones. However, the task of attributing a generated image to its originating source remains relatively unexplored and inherently complex. With the current level of realism achieved by modern generative models, traditional methods reliant on human inspection for attribution have become impractical. While identifying whether an image was generated by a specific model may seem straightforward, it presents nuanced challenges. A simplistic approach involves training a classifier on a dataset comprising both real and generated images produced by the model in question. However, such an approach is susceptible to dataset bias [31] and may struggle to generalize effectively when applied to new data. Furthermore, detectors tailored to specific generative models risk obsolescence as generation techniques evolve and the model they were trained on becomes outdated.

Pre-trained large vision-language models have recently emerged as a promising solution for a multitude of natural language processing and computer vision tasks. These models undergo training on vast image-text datasets sourced from the Internet and exhibit proficiency as zero-shot and few-shot learners for downstream tasks, particularly in applications like image classification [36], detection [22], and segmentation [38]. Moreover, there has been a recent surge in leveraging these models for the detection of synthetic images [4,8,15].

In the current state-of-art, the detection and attribution of synthetic images often face significant challenges. One of the main difficulties lies in the fact that these tasks are typically handled separately, which can lead to ineffective and less robust solutions. Multi-level or cascade architectures are commonly proposed to address these tasks, but they introduce complexity and can be difficult to generalize across different types of synthetic images. The separation of detection and attribution tasks overlooks the potential synergies that could be leveraged by treating them as related tasks. Additionally, the generalization capabilities of existing models are often limited, which hampers their effectiveness in handling diverse and evolving state-of-the-art image generation techniques.

To address these challenges, we introduce FIDAVL, a novel and efficient multitask method that combines synthetic image detection and attribution within a unified framework. Leveraging a vision-language approach, FIDAVL harnesses synergies between vision and language models along with a soft adaptation strategy. This integration enables precise detection and accurate attribution of generated images to their original source models, capitalizing on shared features between the two tasks. Our approach benefits from the generalization capabilities of VLMs, which represents a significant advancement over traditional methods. By treating synthetic image detection and attribution as related tasks within a single-step process, FIDAVL overcomes the limitations of multi-level or cascaded architectures. Extensive experiments conducted on a large-scale dataset including synthetic images generated by various state-of-the-art models demonstrate the high accuracy and robustness of FIDAVL. This approach not only simplifies the process of detection and attribution but also enhances its reliability and scalability. To the best of our knowledge, this study pioneers the utilization of vision-language models for synthetic image attribution and detection in a unified framework.

Our contributions to this paper can be summarized as follows:

- We introduce FIDAVL, a novel single-step approach for synthetic image detection and attribution. Leveraging the complementarity between vision and language, FIDAVL effectively detects and attributes synthetic images to their respective source generation models.
- We adopt a soft prompt-tuning technique to refine the query of FIDAVL for optimal effectiveness.

Through extensive evaluation on a large-scale dataset, our proposed approach demonstrates competitive performance, underscoring its effectiveness in synthetic image detection and attribution. FIDAVL achieves an average accuracy (ACC) exceeding 95% in the synthetic image detection task, and yielding an average ROUGE-L score of 96.50% and an F1-score of 92.64% in the synthetic image attribution task.

The remainder of this paper is organized as follows. Section 2 provides a brief review of the background and related work. Section 3 describes the proposed FIDAVL approach for the attribution and detection of synthetic images. Then, the performance of the proposed approach is assessed and analysed in Section 4. Finally, Section 5 concludes the paper.

2 Background and Related Work

In this section, we delve into generative models, examine advanced deepfake detection and attribution techniques, and offer insights into vision-language models and prompt tuning.

2.1 Generative Models

Generative models have emerged as powerful tools for synthesizing realistic data across various modalities, including images, text, videos, and intricate structures. These models, often harnessed through neural networks, adeptly learn to capture and replicate the underlying patterns and distributions inherent in the training data [10]. Within the domain of deep generative models, a prominent category is generative adversarial network (GAN) [11]. More recently, diffusion models [30] have gained traction as a de-facto method for image generation. The extension of such models to text-to-image synthesis [23,26] has ushered in a wave of models characterized by remarkable quality and diversity, exemplified by models like Imagen [27] and DALL-E-2 [24]. However, the proliferation of deep generative models in image synthesis has also given rise to challenges pertaining to synthetic image detection and attribution.

2.2 Synthetic Image Detection and Attribution

Recent strides in generative models, particularly diffusion-based architectures and cutting-edge GAN models, present challenges to existing detection methodologies. Research highlighted in [7,25] underscores the struggle of current detectors to adapt to these innovative models, underscoring the need for more effective detection techniques. Consequently, a spectrum of novel approaches has emerged in response. Coccomini *et al.* [6] experiment with multi-layer perceptrons (MLPs) and conventional convolutional neural networks (CNNs), probing their efficacy in this domain. Conversely, Wang *et al.*[33] introduce DIRE, a method tailored for diffusion-generated images, which prioritizes the analysis of reconstruction errors. Leveraging diffusion patterns, SeDID [21] achieves accurate detection, with a focus on reverse and denoising computation errors. Amoroso *et al.* [2] explore semantic-style disentanglement to bolster stylistic detection, while Xi *et al.* [35] propose a dual-stream network that emphasizes texture for artificial intelligence (AI)-generated image detection. Wu *et al.* [34] advocate for language-guided synthesis detection (LASTED), treating detection as an identification problem and leveraging language-guided contrastive learning. Ju *et al.* [14] propose a feature fusion mechanism, combining ResNet50 and attention-based modules, for global and local feature fusion in AI-synthesized image detection. Sinitsa *et al.* [29] introduce a rule-based method harnessing CNNs to extract distinctive features, achieving high accuracy even with limited generative image data. In a departure from traditional approaches, Chang *et al.* [4] draw from VLMs, framing deepfake detection as a visual question-answering task. Finally, Cozzolino *et al.*[8] propose a lightweight strategy based on contrastive language image pretraining (CLIP) features and linear support vector machine (SVM), showcasing an alternative avenue for effective detection in this rapidly evolving landscape.

Attributing deepfake content to its source constitutes a crucial aspect in the realm of detection and prevention. Unlike conventional binary detection, attribution introduces a multi-class dimension, facilitating the identification of the specific generative model responsible for the content. Recent studies have shed light on the importance of enhancing attribution techniques. He *et al.*[13] extended detectors to explore textual attribution, revealing areas ripe for improvement in this domain. In the realm of generative visual data, attribution methodologies tailored for GANs have emerged. Bui *et al.* [3] introduced a GAN-fingerprinting technique, which notably enhances source attribution in a closed-set scenario. Recent advancements have also focused on diffusion models (diffusion models (DMs)). Sha *et al.* [28] utilized ResNet for detecting and attributing synthetic images to their respective generators, while Guarnera *et al.* [12] proposed a multi-level approach for synthetic image detection and attribution. Lorenz *et al.* [20] introduced multiLID, a method tailored for diffusion-generated image detection and attribution, leveraging intrinsic dimensionality for enhanced accuracy. Moreover, Wang *et al.* [32] addressed the attribution of generative data to their training data counterparts, necessitating the identification of significant contributors within the training set.

2.3 Vision Language Models

Recent advancements in VLMs have addressed limitations inherent in earlier models, particularly in terms of task specificity and dataset constraints. Noteworthy models such as CLIP, trained on an extensive dataset comprising 400 million image-caption pairs, exemplify this progress by featuring both image and text encoders, thereby facilitating versatile image classification tasks. Leading the charge in this domain are pioneering models such as LLaVA [18], BLIP2 [17], InstructBLIP [9], and Flamingo [1], which represent the vanguard of VLMs innovation. LLaVA, an open-source endeavor, seamlessly integrates vision and language understanding within a vast multimodal framework. BLIP2, on the other hand, achieves state-of-the-art performance through the integration of pre-trained image encoders and language models. Building upon BLIP2, InstructBLIP refines its architecture further, specifically tailoring it for visual instruction tuning. Notably, Flamingo, a family of VLMs, stands out for its adeptness in handling interleaved visual and textual data, thereby making significant strides in adapting to downstream tasks and expanding zero-shot capabilities. These advancements mark a significant leap forward in the realm of VLMs, showcasing their potential to revolutionize various domains reliant on multimodal understanding and processing.

2.4 Prompt Tuning for Vision Language Models

VLMs excel in learning from multimodal data, yet encounter challenges when tasked with adapting to specific downstream vision-related objectives. Groundbreaking research by [37] introduced context optimization (CoOp) to augment the efficiency of CLIP in image classification tasks. Diverging from conventional prompt templates, CoOp learns prompt embeddings with minimal reliance on downstream dataset samples. Prompt tuning manifests in two primary forms: hard and soft. Hard prompt tuning, as proposed in [39], involves adjusting non-differentiable tokens to align with user-defined criteria, albeit encountering difficulties in achieving discrete improvements. Conversely, soft prompt tuning, showcased by [16], optimizes a trainable tensor through back-propagation, thereby enhancing modeling performance. In a notable application, [5] employed subtle prompt optimization techniques to enhance instruction generation in a black-box machine learning (ML) model. These endeavors underscore the importance of nuanced prompt tuning methodologies in enhancing the adaptability and performance of vision-language models across various downstream tasks.

3 Proposed Synthetic Image Detection and Localization

3.1 Problem Formulation

To harness the capabilities of a vision-language model, such as InstructBLIP, we have embraced a framework known as visual question answering (VQA), which we refer to as FIDAVL. FIDAVL is meticulously crafted to respond to inquiries regarding a given image. The input comprises two crucial components: a query image, denoted as I, which serves as the focal point of our scrutiny, and a composite question, denoted as q, which guides FIDAVL in its analysis of the query image. Subsequently, the image is classified as either real or fake; if fake, it is then attributed to its source. The question q can take on various forms, ranging from predefined inquiries like ”Is this photo fake, and what is its source generator?” to customizable questions incorporating a pseudo-word S^*. This adaptability empowers us to tailor our questioning strategy to the specific requirements of our investigation.

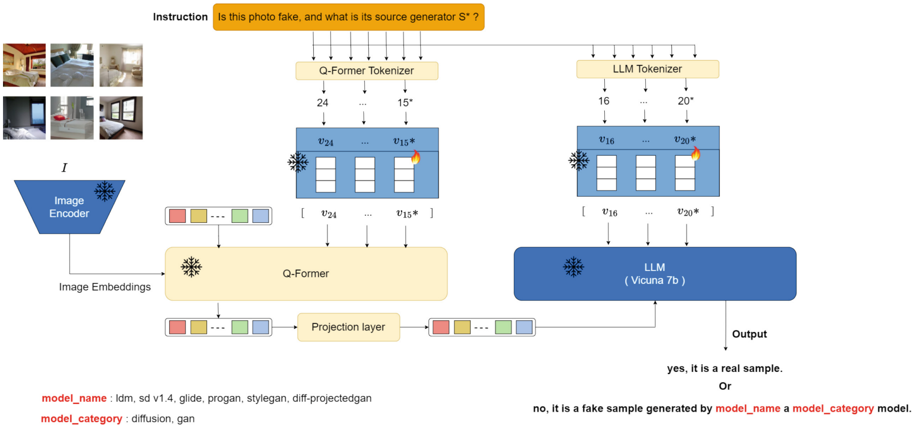

Instruction
Is this photo fake, and what is its source generator S* ?
Q-Former Tokenizer
LLM Tokenizer
24
...
15*
16
...
20*
v_{24}
$v_{15}*$
v_{16}
$v_{20}*$
[v_{24} ... $v_{15}*$]
[v_{16} ... $v_{20}*$]
I
Image Encoder
Image Embeddings
Q-Former
LLM (Vicuna 7b)
Projection layer
Output
yes, it is a real sample.
Or
no, it is a fake sample generated by model_name a model_category model.
model_name : ldm, sd v1.4, glide, progan, stylegan, diff-projectedgan
model_category : diffusion, gan

Fig. 1. Architecture of the proposed synthetic image detection and localization.

The output of FIDAVL comprises a set of response texts, denoted as $\hat{y}$. While $\hat{y}$ theoretically encompasses any text, we impose specific constraints to uphold consistency and clarity in our responses. If the query image is determined to be real, the response is articulated as **”No, it is a real sample.”**. Conversely, if it is deemed fake, the response adheres to the template **”Yes, it is a fake sample generated by** *model_name*, a *model_category***model.”**. Here, *model_name* signifies the name of the generating model, which could belong to the set progan, diff-projectedgan, stylegan, ldm, glide, Stable diffusion, while *model_category* denotes the category of the generating model, which could be diffusion or gan. This response structure aligns with our ground truth for synthetic image detection and attribution. Finally, to evaluate the efficacy of FIDAVL, we gauge the accuracy of both the detection and attribution tasks. This quantitative assessment offers insights into our model's proficiency in accurately identifying and attributing synthetic images.

Mathematically, the formulation of the single-step synthetic image detection and attribution task is as follows:

$$\hat{y} = \mathcal{M}_{\theta}(I, q). \tag{1}$$

where $\mathcal{M}$ is an VLM with parameters θ, which takes an image I and a question q as input and generates an answer $\hat{y}$.

3.2 Soft Prompt Tuning

Our investigation harnesses soft prompt tuning within InstructBLIP, following the outlined procedure. In InstructBLIP, the prompt serves as input to two pivotal components: Q-Former and large language model (LLM). Initially, the prompt undergoes tokenization and embedding before being concurrently fed into both Q-Former and the LLM, as illustrated in Fig. 1. To facilitate prompt tuning, we introduce a pseudo-word S^* into the prompt, which acts as the target for tuning. Specifically, we adopt the question pattern "Is this photo fake, and what is its source generator?", appending the pseudo-word to the end of the prompt. This modification yields the following adjusted prompt q^*: "Is this photo fake, and what is its source generator S^*?". For real images, we assign the output label y as "No, it is a real sample." Conversely, for fake images, the label y is set as "Yes, it is a fake sample generated by *model_name*, a *model_category* model." This labeling scheme facilitates soft prompt tuning.

We then proceed to freeze all model modules except the word embedding v^* corresponding to the pseudo-word S^*, which is randomly initialized. Subsequently, we optimize the word embedding v^* of the pseudo-word across a triplet training set $\{I, q^*, y\}$ using the language modeling loss. Our aim is to align the output of the VLM, denoted as $\hat{y}$, with the label y. Our optimization objective can therefore be defined as :

$$f_{S^*} = \arg\min_{S^*} \mathbb{E}_{(I,y)} \left[L(M(I, q^*), y) \right] \tag{2}$$

where L is the language modeling loss function (cross-entropy loss).

4 Experimental Results

Dataset. The dataset utilized in this study is a meticulously curated collection of images comprising two primary components: real images sourced from the largescale scene understanding (LSUN) bedroom dataset and synthetic data generated by three distinct GAN engines (ProGAN, StyleGAN, Diff-ProjectedGAN), as well as three text-to-image DM models (LDM, Glide, Stable diffusion v1.4). For each considered GAN, 20,000 images were generated for training and an additional 10,000 for testing, resulting in a total of 90,000 synthetic images. Similarly, each DM architecture generated an equivalent number of images for both training and testing, leveraging the prompt "A photo of a bedroom", thus yielding another 90,000 images. Consequently, the cumulative synthetic dataset comprises 180,000 images. In addition to synthetic data, the dataset incorporates 130,000 real images. Notably, the real images designated for testing remain consistent across all testing subsets.

Implementation Details. We use the GitHub repository of [4] based on LAVIS library for implementation, training, and evaluation. To prevent out-of-memory issues on small GPU, we employ Vicuna-7B as LLM. For prompt tuning, we initialize the model with an instruction-tuned checkpoint from LAVIS, exclusively fine-tuning the word embeddings of the pseudo-word while freezing the rest of the model. The model is prompt-tuned with a maximum of 5 epochs, employing the AdamW optimizer with $\beta_1 = 0.9$ and $\beta_2 = 0.999$, batch size 16, and a weight decay of 0.05. The initial learning rate is set to 10^{-8}, and apply cosine decay with a minimum learning rate of 0. The code is executed on an NVIDIA RTX A4500 GPU with 16 GB and an Intel(R) i9-12950HX CPU with Windows 11 Pro. In terms of image processing, all the images are resized to 224 pixels on the shorter side, maintaining the original aspect ratio. In training, random cropping yields a final size of 224×224 pixels, while testing involves center cropping to the same size.

Evaluation Metrics. In our synthetic image detection and attribution task, we evaluate our FIDAVL model across multiple metrics including accuracy, F1-score. Since we cannot directly compare results from textual data as if it were binary classification, what we can do is calculate overlapping words between predictions and references. In this regard, we use the ROUGE score, which measures the degree of correspondence between the content of the generated sentence and the content of a set of reference sentences. The higher the value of these metrics, the better the performance of the model.

4.1 Synthetic Image Detection

In this section, we delve into an extensive analysis of these results, meticulously examining the model's performance across our test set and elucidating the strengths of our detection strategy. Through a comprehensive examination of metrics such as accuracy (ACC) and F1 score, we aim to gain deeper insights into the efficacy with which FIDAVL tackles the task of synthetic image detection.

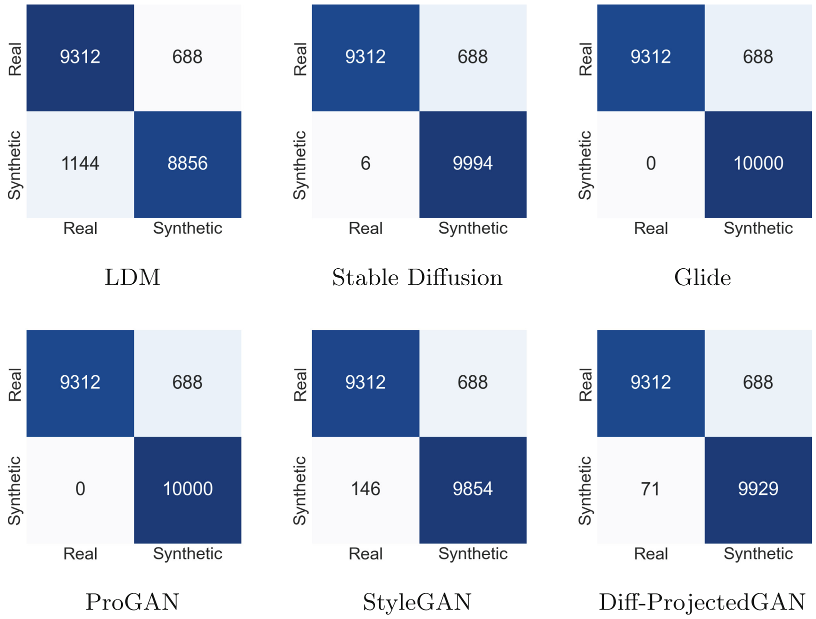
Table 1 showcases the evaluation outcomes concerning the detection capabilities of our proposed method, FIDAVL. Across all test subsets, FIDAVL showcased robust performance, consistently attaining high accuracy and F1 scores. Remarkably, FIDAVL achieved an average accuracy of **95.42%** alongside an impressive F1 score of **95.47%**, underscoring its effectiveness in precisely distinguishing between synthetic and authentic images.

Table 1. Synthetic image detection task and comparison to baseline models. We report ACC (%) / F1-Score (%). Note that, on average (two last columns), our model yields better performance.

Method	Testing Subset						Average(in %)
	LDM*	SD v1.4*	GLIDE*	ProGAN $^\oplus$	StyleGAN$^\oplus$	Diff-ProjectedGAN$^\oplus$	
ResNet50	99.92 / 99.92	75.47 / 67.57	73.10 / 63.28	94.28 / 93.94	77.94 / 71.75	59.20 / 31.27	79.98 / 71.29
Xception	**99.96 / 99.96**	63.84 / 43.41	58.92 / 30.35	64.50 / 45.11	69.96 / 57.18	51.14 / 04.79	68.05 / 46.80
DeiT	99.83 / 99.83	96.02 / 95.86	**98.15 / 98.11**	93.28 / 92.81	95.08 / 94.84	77.06 / 70.32	93.23 / 91.96
FIDAVL	90.84 / 90.62	**96.53 / 96.64**	96.56 / 96.67	**96.56 / 96.67**	**95.83 / 95.94**	**96.20 / 96.31**	**95.42 / 95.47**

* Diffusion-based model. $^\oplus$ GAN-based model.

The efficacy of FIDAVL can be attributed to its innovative approach, leveraging the complementary strengths inherent in vision and language modalities. By seamlessly integrating both vision and language models, FIDAVL harnesses the semantic understanding embedded within each modality, enabling it to discern nuanced cues and patterns indicative of synthetic image generation. This underscores the significance of interdisciplinary methodologies in crafting resilient solutions to intricate challenges like synthetic image detection.



LDM Stable Diffusion Glide

ProGAN StyleGAN Diff-ProjectedGAN

Fig. 2. Confusion matrices per testing subset on synthetic image detection task.

Fig. 2 provides a comprehensive overview of FIDAVL's performance in differentiating synthetic image samples from real ones. Each subfigure depicts a confusion matrix corresponding to a specific testing subset, labeled accordingly.

Across all subsets, a consistent false negative rate of 688 is observed, underscoring a shared challenge in accurately detecting synthetic images. Notably, the most promising results are observed in the glide and progan subsets, where all synthetic images were detected. However, FIDAVL encounters challenges in accurately detecting LDM-generated images, as evidenced by a significant number of true positives, totaling 1144. This difficulty can be attributed to the homogeneity of our specific bedroom image dataset, which presents distinct characteristics that may pose challenges for detection algorithms.

Fig. 3 provides an in-depth analysis of the distribution of well-detected synthetic images according to whether they were generated by GAN-based or diffusion-based models. In Fig. 2, we observed from the LDM confusion matrix that 8856 synthetic images were well detected. Furthermore, in Fig. 3, the LDM confusion matrix illustrates the distribution of these images based on their attribution to the respective generator source model type, 8266 to diffusion and 590 to GAN. Fig. 3 shows that although the images have been well classified as synthetic, FIDAVL encounters challenges in accurately attributing these images to their specific source model type, a phenomenon particularly observed with GAN-based test sets and LDM. Moreover, the best performances are obtained on stable diffusion and glide.

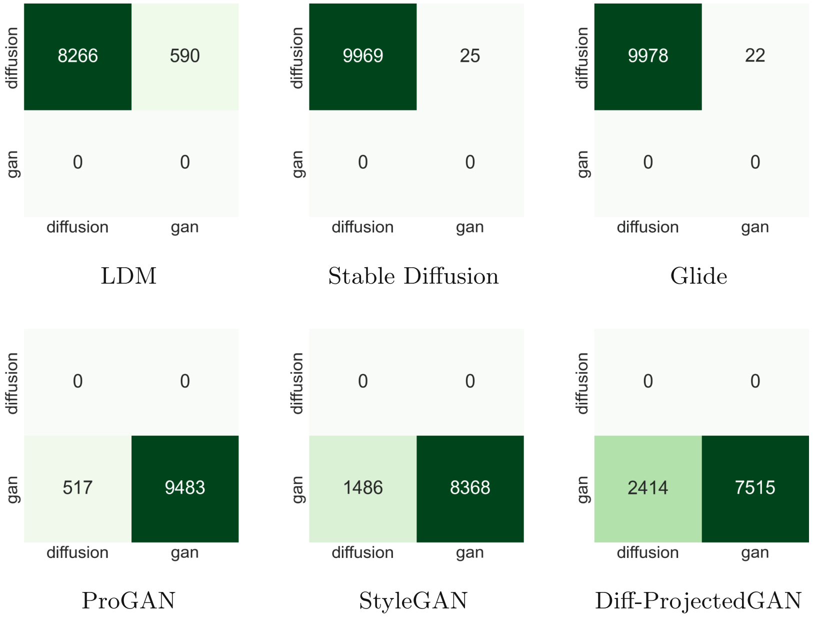


Fig. 3. Confusion matrices indicate which synthetic images detected as synthetic are correctly classified according to their generating source model.

Comparative analysis. In this subsection, we conduct a comparative analysis of FIDAVL against three baseline models: ResNet50, Xception, and DeiT. To establish our baseline models, we fine-tuned these architectures by replacing their final FC layers with a novel FC layer containing a single neuron dedicated to distinguishing real images from fake ones. These models were initialized with pre-trained weights obtained from the ImageNet dataset, thereby leveraging the knowledge encoded in their learned representations. We evaluate each model's performance across multiple testing subsets, including LDM, SD v1.4, GLIDE, ProGAN, StyleGAN, and Diff-ProjectedGAN. We present the average performance across these subsets to offer a comprehensive view of the models' effectiveness.

Table 1 summarized the obtained results from the experiment. ResNet50 performs exceptionally well, particularly in the LDM subset with 99.92% accuracy and 99.92% F1 score, and maintains good performance across other subsets with an average accuracy of 79.98% and F1 score of 71.29%. Xception shows comparable accuracy in the LDM (99.96%), but declines considerably in the other subsets, with an average accuracy of 68.05% and an F1 score of 46.80%. DeiT demonstrates strong performance, especially in the SD v1.4 (96.02% accuracy and 95.86% F1 score) and GLIDE (98.15% accuracy and 98.11% F1 score) subsets, with an average accuracy of 93.23% and an F1 score of 91.96%. In contrast, FIDAVL exhibits outstanding performance across all subsets, with an average accuracy of 95.42% and an F1 score of 95.47%. In particular, FIDAVL excels in SD v1.4, ProGAN, StyleGAN, and Diff-ProjectedGAN subsets, showcasing its robustness and competitiveness compared to the baseline models.

To summarize, our approach shows competitive performance, albeit with lower scores in testing subsets such as LDM and GLIDE. Notably, FIDAVL reaches around 90.84% on LDM and maintains scores above 95% on other subsets. FIDAVL adopts a multitask learning approach, which not only involves image detection (distinguishing real from fake) but also includes an attribution task aimed at identifying the model responsible for generating a given image. This dual-focus training introduces additional complexity and objectives to the model's training regimen, which can likely influence its performance dynamics as it must balance learning across multiple objectives.

Generalization to unseen generative models. In this subsection, we evaluate FIDAVL generalization capabilities on multiple unseen synthetic image detection subsets, including ADM, DDPM, IDDPM, PNDM, Diff-StyleGAN2, and ProjectedGAN. Each subset represents distinct characteristics and challenges within the detection task, enabling a comprehensive assessment of FIDAVL's generalization capabilities.

Results in Table 2 highlight FIDAVL's generalization performance across the different subsets. Overall, FIDAVL generalizes very well, with an average accuracy of 86.04% and F1-score of 83.48% across all unseen test sets during training.

Table 2. Generalization results on synthetic images generated by unseen generation models. We report ACC (%) / F1-Score (%).

Method	Testing Subsets						Average(in %)
	ADM*	DDPM*	IDDPM*	PNDM*	Diff-StyleGAN2$^\oplus$	ProjectedGAN$^\oplus$	
ResNet50	**72.32 / 61.82**	75.26 / 67.21	**88.96 / 87.61**	77.20 / 70.52	61.62 / 37.88	58.35 / 28.82	72.28 / 58.98
Xception	52.05 / 07.98	58.60 / 29.41	54.62 / 16.99	60.01 / 33.43	71.53 / 60.03	51.64 / 06.66	58.08 / 25.75
DeiT	50.40 / 02.01	50.18 / 01.17	50.14 / 01.01	56.25 / 22.54	93.26 / 92.79	79.84 / 74.82	63.34 / 32.39
FIDAVL	67.35 / 56.01	**86.56 / 85.61**	81.38 / 78.91	**94.93 / 95.02**	**96.25/ 96.36**	**89.78 / 88.98**	**86.04 / 83.48**

* Diffusion-based model. $^\oplus$ GAN-based model.

ResNet50 demonstrates moderate performance across subsets, showing notable strength in ADM and IDDPM, while Xception exhibits variable performance, particularly struggling with ADM, DDPM, and IDDPM subsets. DeiT performs similarly to Xception, facing challenges in ADM, DDPM, and IDDPM subsets. FIDAVL shows superior performance across most subsets, especially excelling in DDPM, IDDPM, PNDM, and GAN-based subsets like Diff-StyleGAN2 and ProjectedGAN.

Moreover, the results reveal patterns and considerations that need further investigation:

- ADM* subset: FIDAVL achieves an accuracy of 67.35% and F1-score of 56.01%, indicating moderate performance.
- DDPM* subset: Fake Image Detect and Attribution using a Vision-Language model (FIDAVL) achieved a commendable accuracy of 86.56% and an F1-score of 85.61%, suggesting strong performance in detecting diffusion-based models. However, deeper analysis is warranted to understand any potential biases or limitations when handling these types of synthetic images.
- IDDPM* subset: FIDAVL's performance (accuracy: 81.38%, F1-score: 78.91%) indicates slightly reduced effectiveness compared to other subsets, suggesting potential challenges in detecting specific characteristics associated with this subset, and necessitating further investigation into the model's adaptability.
- PNDM* subset: FIDAVL excelled with an impressive accuracy of 94.93% and an F1-score of 95.02%, indicating robust performance in detecting certain types of diffusion-based models. Besides, this highlights its strengths but raises questions about its generalizability across all diffusion-based variants.
- Diff-StyleGAN2$^\oplus$ subset: FIDAVL demonstrated high accuracy (96.25%) and a high F1-score (96.36%) in detecting this GAN-based model. Although this achievement underlines the ability of FIDAVL to identify this specific GAN architecture, further research is needed to assess its performance over a wider range of GAN variations.
- ProjectedGAN$^\oplus$ subset: FIDAVL demonstrates strong performance with an accuracy of 96.38% and an f1-score of 96.49%. This showcases FIDAVL's ability to accurately detect images generated by ProjectedGAN models.

Although FIDAVL shows promising performance, a rather critical aspect deserves closer investigation. FIDAVL's exceptional performance on certain subsets raises questions about its focus on specific model characteristics versus broader synthetic image detection. However, the balance between model specificity and general applicability is essential for its deployment in the real world. The results underline FIDAVL's effectiveness in handling diverse synthetic image datasets generated by unseen models. Its superior performance signifies strong generalization potential, critical for real-world applications where model adaptability to varying synthetic data sources is essential.

4.2 Synthetic Image Attribution

In this section, we assess the performance of FIDAVL in the synthetic image attribution task using ROUGE scores as metrics, in conjunction with standard classification metrics such as accuracy and F1-score. As detailed in Subsection 3.1, FIDAVL generates text as output. ROUGE scores are widely recognized as metrics commonly used in text generation tasks. These scores primarily gauge the quality of machine-generated text by comparing it to reference text, measuring various aspects of text similarity, such as overlap in n-grams (consecutive sequences of words). Furthermore, the inclusion of accuracy and F1-score provides a comprehensive understanding of FIDAVL's performance in synthetic image attribution. In our experiment, we utilize two ROUGE scores: ROUGE-2 and ROUGE-L.

Table 3. Performance evaluation of synthetic image attribution task.

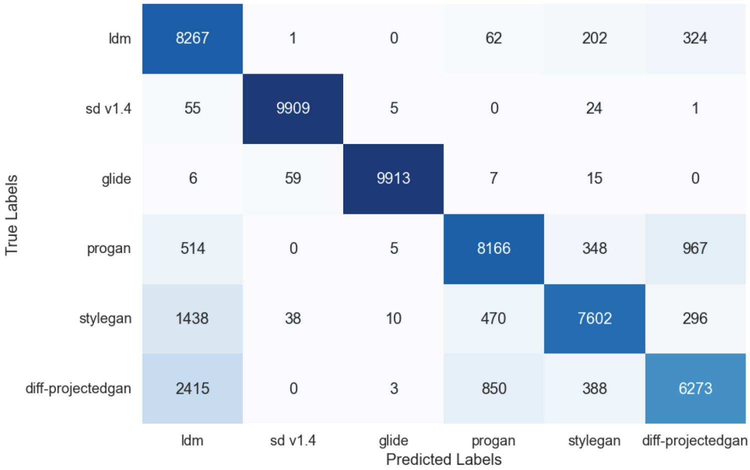
Method	ROUGE-2 / ROUGE-L scores on different testing subsets						Average (in %)
	LDM*	SD v1.4*	GLIDE*	ProGAN$^{\oplus}$	StyleGAN$^{\oplus}$	Diff-ProjectedGAN$^{\oplus}$	
FIDAVL	92.23 / 94.82	97.39 / 98.19	**97.41 / 98.20**	94.99 / 97.01	93.21 / 96.14	90.62 / 94.64	94.30 / 96.50
Method	ACC / F1-score on different testing subsets						Average (in %)
	LDM*	SD v1.4*	GLIDE*	ProGAN$^{\oplus}$	StyleGAN$^{\oplus}$	Diff-ProjectedGAN$^{\oplus}$	
FIDAVL	87.89 / 89.27	96.10 / 97.96	**96.12 / 98.00**	87.39 / 93.17	84.57 / 90.95	77.92 / 86.54	88.33 / 92.64

* Diffusion-based model. $^{\oplus}$ GAN-based model.

Table 3 presents a comprehensive evaluation of FIDAVL in synthetic image attribution task across different test sets classified according to their underlying architectures: diffusion models (LDM, Stable Diffusion v1.4, GLIDE) and GAN models (ProGAN, StyleGAN, Diff-ProjectedGAN). The evaluation metrics used are ROUGE-2, ROUGE-L, accuracy, and F1-score, measured on different test subsets.

First, the results show that FIDAVL generally achieves competitive performance in terms of ROUGE scores, accuracy, and F1-score on diffusion-based models compared to GAN-based models. In particular, Stable Diffusion v1.4 and GLIDE achieve higher ROUGE scores, accuracy and F1-score than ProGAN, StyleGAN, and Diff-ProjectedGAN. This variation highlights the sensitivity of FIDAVL to the characteristics inherent in different architectural models, potentially indicating the model's proficiency in specific image generation paradigms.

Fig. 4 illustrates the distribution of accurately classified synthetic images across various generative models. The diagonal elements (True Positive) depict the number of correct predictions for each category. Remarkably, FIDAVL demonstrates exceptional performance on stable diffusion and Glide, with 9909 and 9913 instances correctly classified, respectively. However, the matrix also sheds light on areas of concern. FIDAVL encounters difficulties in accurately attributing GAN-based generated images to their specific source models. Many GAN-based generated images are incorrectly attributed to LDM and other GAN-based models. This may be attributed to the fact that unconditional diffusion models, such as LDM, share similarities with GAN-based generative models, posing challenges for accurate attribution.



	ldm	sd v1.4	glide	progan	stylegan	diff-projectedgan
ldm	8267	1	0	62	202	324
sd v1.4	55	9909	5	0	24	1
glide	6	59	9913	7	15	0
progan	514	0	5	8166	348	967
stylegan	1438	38	10	470	7602	296
diff-projectedgan	2415	0	3	850	388	6273

Fig. 4. Confusion Matrix for Attribution Task: Synthetic data correctly classified as synthetic but attributed to a different source from the generating source.

5 Conclusion and Future Work

In this paper, we have proposed FIDAVL, a novel multitask framework for AI-generated image detection and attribution leveraging vision-language models. Through the integration of vision and language modalities, FIDAVL exhibited exceptional performance in accurately discerning and attributing AI-generated images to their respective source models. Extensive experimentation validated the effectiveness of FIDAVL in addressing the challenges of synthetic image detection and attribution simultaneously. Our findings underlined the significance of interdisciplinary approaches in tackling complex problems in today's rapidly evolving technological landscape. With its promising performance, FIDAVL presented a valuable solution to enhance accountability and trust amidst the proliferation of fake images. In future endeavors, we aim to conduct additional experiments to evaluate the robustness and generalization capabilities of FIDAVL in real-world scenarios. This includes exploring scenarios involving JPEG compression, scaling, unseen images from new generative models, and added noise. Additionally, we plan to extend FIDAVL into a multi-head vision-language framework to further enhance its capabilities and versatility.

Acknowledgments. This work has been partially funded by the project PCI2022-134990-2 (MARTINI) of the CHISTERA IV Cofund 2021 program. Abdenour Hadid is funded by TotalEnergies collaboration agreement with Sorbonne University Abu Dhabi.

References

1. Alayrac, J.B., Donahue, J., Luc, P., Miech, A., Barr, I., Hasson, Y., Lenc, K., Mensch, A., et al.: Flamingo: a visual language model for few-shot learning. Adv. Neural. Inf. Process. Syst. **35**, 23716–23736 (2022)
2. Amoroso, R., Morelli, D., Cornia, M., Baraldi, L., Del Bimbo, A., Cucchiara, R.: Parents and children: Distinguishing multimodal deepfakes from natural images. arXiv preprint arXiv:2304.00500 (2023)
3. Bui, T., Yu, N., Collomosse, J.: Repmix: Representation mixing for robust attribution of synthesized images. In: European Conference on Computer Vision. pp. 146–163. Springer (2022)
4. Chang, Y.M., Yeh, C., Chiu, W.C., Yu, N.: Antifakeprompt: Prompt-tuned vision-language models are fake image detectors. arXiv preprint arXiv:2310.17419 (2023)
5. Chen, L., Chen, J., Goldstein, T., Huang, H., Zhou, T.: Instructzero: Efficient instruction optimization for black-box large language models. arXiv preprint arXiv:2306.03082 (2023)
6. Coccomini, D.A., Esuli, A., Falchi, F., Gennaro, C., Amato, G.: Detecting images generated by diffusers. arXiv preprint arXiv:2303.05275 (2023)
7. Corvi, R., Cozzolino, D., Zingarini, G., Poggi, G., Nagano, K., Verdoliva, L.: On the detection of synthetic images generated by diffusion models. In: ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 1–5. IEEE (2023)
8. Cozzolino, D., Poggi, G., Corvi, R., Nießner, M., Verdoliva, L.: Raising the bar of ai-generated image detection with clip. arXiv preprint arXiv:2312.00195 (2023)
9. Dai, W., Li, J., Li, D., Tiong, A., Zhao, J., Wang, W., Li, B., Fung, P., Hoi, S.: Instructblip: Towards general-purpose vision-language models with instruction tuning. arxiv 2023. arXiv preprint arXiv:2305.06500 **2** (2023)
10. Esser, P., Rombach, R., Ommer, B.: Taming transformers for high-resolution image synthesis. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 12873–12883 (2021)
11. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. Advances in neural information processing systems **27** (2014)
12. Guarnera, L., Giudice, O., Battiato, S.: Level up the deepfake detection: a method to effectively discriminate images generated by gan architectures and diffusion models. arXiv preprint arXiv:2303.00608 (2023)
13. He, X., Shen, X., Chen, Z., Backes, M., Zhang, Y.: Mgtbench: Benchmarking machine-generated text detection. arXiv preprint arXiv:2303.14822 (2023)

14. Ju, Y., Jia, S., Cai, J., Guan, H., Lyu, S.: Glff: Global and local feature fusion for ai-synthesized image detection. IEEE Transactions on Multimedia (2023)
15. Keita, M., Hamidouche, W., Bougueffa Eutamene, H., Hadid, A., Taleb-Ahmed, A.: Bi-lora: A vision-language approach for synthetic image detection. ArXiv (2024)
16. Lester, B., Al-Rfou, R., Constant, N.: The power of scale for parameter-efficient prompt tuning. arXiv preprint arXiv:2104.08691 (2021)
17. Li, J., Li, D., Savarese, S., Hoi, S.: Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. arXiv preprint arXiv:2301.12597 (2023)
18. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning. arXiv preprint arXiv:2304.08485 (2023)
19. Liz-López, H., Keita, M., Taleb-Ahmed, A., Hadid, A., Huertas-Tato, J., Camacho, D.: Generation and detection of manipulated multimodal audiovisual content: Advances, trends and open challenges. Information Fusion **103**, 102103 (2024)
20. Lorenz, P., Durall, R., Keuper, J.: Detecting images generated by deep diffusion models using their local intrinsic dimensionality. preprint arXiv:2307.02347 (2023)
21. Ma, R., Duan, J., Kong, F., Shi, X., Xu, K.: Exposing the fake: Effective diffusion-generated images detection. arXiv preprint arXiv:2307.06272 (2023)
22. Ming, Y., Cai, Z., Gu, J., Sun, Y., Li, W., Li, Y.: Delving into out-of-distribution detection with vision-language representations. Adv. Neural. Inf. Process. Syst. **35**, 35087–35102 (2022)
23. Nichol, A., Dhariwal, P., Ramesh, A., Shyam, P., Mishkin, P., McGrew, B., Sutskever, I., Chen, M.: Glide: Towards photorealistic image generation and editing with text-guided diffusion models. arXiv preprint arXiv:2112.10741 (2021)
24. Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., Chen, M.: Hierarchical text-conditional image generation with clip latents. ArXiv:2204.06125 **1**(2), 3 (2022)
25. Ricker, J., Damm, S., Holz, T., Fischer, A.: Towards the detection of diffusion model deepfakes. arXiv preprint arXiv:2210.14571 (2022)
26. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10684–10695 (2022)
27. Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E.L., Ghasemipour, K., Gontijo Lopes, R., Karagol Ayan, B., Salimans, T., et al.: Photorealistic text-to-image diffusion models with deep language understanding. Adv. Neural. Inf. Process. Syst. **35**, 36479–36494 (2022)
28. Sha, Z., Li, Z., Yu, N., Zhang, Y.: De-fake: Detection and attribution of fake images generated by text-to-image diffusion models. preprint arXiv:2210.06998 (2022)
29. Sinitsa, S., Fried, O.: Deep image fingerprint: Accurate and low budget synthetic image detector. arXiv preprint arXiv:2303.10762 (2023)
30. Song, J., Meng, C., Ermon, S.: Denoising diffusion implicit models. arXiv preprint arXiv:2010.02502 (2020)
31. Torralba, A., Efros, A.A.: Unbiased look at dataset bias. In: CVPR 2011. pp. 1521–1528. IEEE (2011)
32. Wang, S.Y., Efros, A.A., Zhu, J.Y., Zhang, R.: Evaluating data attribution for text-to-image models. arXiv preprint arXiv:2306.09345 (2023)
33. Wang, Z., Bao, J., Zhou, W., Wang, W., Hu, H., Chen, H., Li, H.: Dire for diffusion-generated image detection. arXiv preprint arXiv:2303.09295 (2023)
34. Wu, H., Zhou, J., Zhang, S.: Generalizable synthetic image detection via language-guided contrastive learning. arXiv preprint arXiv:2305.13800 (2023)
35. Xi, Z., Huang, W., Wei, K., Luo, W., Zheng, P.: Ai-generated image detection using a cross-attention enhanced dual-stream network. ArXiv:2306.07005 (2023)

36. Zhang, R., Zhang, W., Fang, R., Gao, P., Li, K., Dai, J., Qiao, Y., Li, H.: Tip-adapter: Training-free adaption of clip for few-shot classification. In: European Conference on Computer Vision. pp. 493–510. Springer (2022)
37. Zhou, K., Yang, J., Loy, C.C., Liu, Z.: Learning to prompt for vision-language models. Int. J. Comput. Vision **130**(9), 2337–2348 (2022)
38. Zhou, Z., Lei, Y., Zhang, B., Liu, L., Liu, Y.: Zegclip: Towards adapting clip for zero-shot semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11175–11185 (2023)
39. Zou, A., Wang, Z., Kolter, J.Z., Fredrikson, M.: Universal and transferable adversarial attacks on aligned language models. arXiv preprint arXiv:2307.15043 (2023)

DeepFeatureX Net: Deep Features eXtractors Based Network for Discriminating Synthetic from Real Images

Orazio Pontorno(✉), Luca Guarnera, and Sebastiano Battiato

Department of Mathematics and Computer Science,
University of Catania, Catania, Italy
orazio.pontorno@phd.unict.it,
{luca.guarnera,sebastiano.battiato}@unict.it

Abstract. Deepfakes, synthetic images generated by deep learning algorithms, represent one of the biggest challenges in the field of Digital Forensics. The scientific community is working to develop approaches that can discriminate the origin of digital images (real or AI-generated). However, these methodologies face the challenge of generalization, that is, the ability to discern the nature of an image even if it is generated by an architecture not seen during training. This usually leads to a drop in performance. In this context, we propose a novel approach based on three blocks called Base Models, each of which is responsible for extracting the discriminative features of a specific image class (Diffusion Model-generated, GAN-generated, or REAL) as it is trained by exploiting deliberately unbalanced datasets. The features extracted from each block are then concatenated and processed to discriminate the origin of the input image. Experimental results showed that this approach not only demonstrates good robust capabilities to JPEG compression and other various attacks but also outperforms state-of-the-art methods in several generalization tests. Code, models and dataset are available at https://github.com/opontorno/block-based_deepfake-detection.

Keywords: Deepfake Detection · Multimedia Forensics · Generative Models

1 Introduction

Generative models have achieved a high degree of fidelity in content generation, producing increasingly realistic and convincing results. Thanks to the vast amount of data available today and the continuous development of complex architectures, such as Generative Adversarial Networks (GANs) [17] and Diffusion Models (DMs) [25,46], these models are able to produce images, text, sound and video with an astonishing quality that can hardly be distinguished

A. Antonacopoulos et al. (Eds.): ICPR 2024, LNCS 15321, pp. 177–193, 2025.
https://doi.org/10.1007/978-3-031-78305-0_12

from those created by human beings. This ability to generate high-fidelity content has opened up new opportunities in a wide range of fields, from art and entertainment to scientific research and multimedia content production. However, along with their powerful creative capabilities, generative models also have several negative aspects. One of the main problems is the possibility of abuse, as such models can be used to generate fake or convincingly manipulated content, fuelling the spread of misinformation and fraud [45,48]. Moreover, they can raise ethical concerns regarding intellectual property and privacy [33], especially when they are used to create content based on personal data without the consent of the involved people. The proper and preventive detection of AI-generated content therefore becomes a critical priority to combat the spread of deepfakes and maintain the integrity of online information.

The scientific community is striving to find increasingly new and effective techniques and methods that can discern the nature (real or generated) of digital images. These techniques can be based on analysis and processing of statistics extracted from images (e.g. analytical traces) or on deep learning engines. Among other we recall the analysis of image frequencies, such as the Discrete Cosine Transform (DCT) and the Fourier Transform to map image pixels from the spatial domain to the frequency domain, facilitating greater interpretability in the task of deepfake recognition [2,22]. Deep learning-based methodologies involve the construction of neural models achieving in general better results than the previous techniques [1,16], but at the expense of a lower generalization.

These approaches showed that GAN-generated and Diffusion Model-generated images have different traces in the Fourier domain. Using a single architecture to classify real and deepfake images can lead to feature contamination. For this reason, in this paper we propose a deep learning based architecture that exploits three backbones, called "Base Models" (BM) trained and specialized to specific classification tasks with special emphasis to DM-generated data, GAN-generated data, and real ones. The fundamental concept is based on utilising the inherent capabilities of the basic models, each of which is dedicated to extracting discriminating features specific to a generating architecture left behind during the image generation process. This approach aims at enhancing the final model by making it more resilient and robust to JPEG compression attacks, commonly employed by social networks, as well as other perturbations such as Gaussian noise, mirroring, rotation, and resizing. The goal is to improve the model's ability to generalize its acquired knowledge and maintain high performance across a range of image modifications. Focusing on specific distinctive features associated with different image generation technologies allows the model to develop a deeper and more focused understanding of the peculiarities of each image category, thus improving its ability to distinguish between genuine and synthetic images in real and variable contexts. With this work we face the difficulty, often encountered in the state-of-the-art, of generalizing the recognition capabilities acquired in the training phase both to images generated by AIs not belonging to the dataset used in that phase and to synthetic images of generating architectures other than those taken into consideration.
The main contributions of this paper are:

- A new approach for extracting main features from digital images using Base Models.
- A model capable of retaining its discriminating capacity even in the presence of different attacks such as JPEG compression, Mirroring, Scaling, and many others.

The paper follows the following structure: Section 2 provides an overview of the main deepfake detection methods currently present in the state-of-the-art; in Section 3, a detailed description of the dataset of images used to conduct the experiments is provided; subsequently, in Section 4.2, the architecture proposed in this study and the stages of the training method are presented in detail; the experimental results obtained during the testing phase are reported in a Section 5.2; finally, the paper concludes with a concluding section where the main findings are summarized and the future directions of research are outlined.

2 Related Works

Most deepfake detection methods are based on intrinsic trace analysis to classify real content from synthetic ones. The Expectation-Maximization algorithm was used in [19] to capture the correlation between pixels, resulting in a discriminative trace able to distinguish deepfake images from pristine ones. McCloskey et al. [38] showed that generative models create synthetic content with color channel curve statistics different from the real data, resulting in another discriminative trace. In the frequency domain [21,37], researchers highlighted the possibility of identifying abnormal traces left during generative models, mainly analyzing features extracted from DCT [3,9,15]. Liu et al. [35] proposed a method called Spatial-Phase Shallow Learning (SPSL) that combines spatial imaging and phase spectrum to capture artifacts from up-sampling on synthetic data, improving deepfake detection. Corvi et al. [11] analyzed a large number of images generated by different families of generative models (GAN, DM, and VQ-GAN (Vector Quantized Generative Adversarial Networks)) in the Fourier domain to discover the most discriminative features between real and synthetic images. The experiments showed that regular anomalous patterns are available in each category of involved architecture. Another category of detectors are deep neural network-based approaches. Wang et al. [50] used a ResNet-50 model trained with images generated by ProGAN [28] to differentiate real from synthesized images. Their study demonstrated the model's ability to generalize beyond ProGAN-generated Deepfakes. Wang et al. [49] introduced FakeSpotter, a new approach that relies on monitoring the behaviors of neurons (counting which and how many activate on the input image) within a dedicated CNN to identify Deepfake-generated faces. Many researchers have focused their research on trying to investigate how possible it is to detect images created by diffusion models. Corvi et al. [10] were among the first to address this issue, exploring the difficulties in distinguishing images generated by diffusion models from real ones and evaluating the suitability of current detectors. Sha et al. [44] proposed DE-FAKE, a machine learning classifier designed for detecting diffusion model-generated images across four prominent text-to-image architectures. The authors then proposed a pioneering study on the detection and attribution of fake images generated by diffusion models, demonstrating the feasibility of distinguishing such images from real ones and attributing them to the source models, and also discovering the influence of prompts on the authenticity of images. Recently, Guarnera et al.[20] proposed a method based on the attribution of images generated by generative adversarial networks (GANs) and diffusion models (DMs) through a multi-level hierarchical strategy. At each level, a distinct and specific task is addressed: the first level (more generic), allows discerning between real and AI-generated images (either created by GAN or DM architectures); the second level determines whether the images come from GAN or DM technologies; and the third level addresses the attribution of the specific model used to generate the images.

The limitations of these methods mainly concern the presence of experimental results performed only under ideal conditions and, consequently, the almost total absence of generalization tests: the classification performance of most state-of-the-art methods drops drastically when testing images generated by architectures never considered during the training procedure.

3 Dataset details

The dataset comprises a total of $72,334$ images, distributed as shown in Table 1.

Table 1. Number, sizes and sources of involved images. The column *Type* specifies whether the image category represents Faces (F) and/or Other (O) (e.g. animals, statues, etc.).

Nature	Architecture	Type	# Images	Total	Different Sizes
GAN	AttGAN [24]	FO	6005	37.572	256 × 256
	BigGAN [4]	O	2600		256 × 256
	CycleGAN [54]	FO	1047		256 × 256; 512 × 512
	GauGAN [40]	O	4000		256 × 256; 512 × 512
	GDWCT [5]	O	3367		216 × 216
	ProGAN [29]	O	1000		256 × 256; 512 × 512
	StarGAN [6]	F	6848		256 × 256
	StyleGAN [31]	O	4705		256 × 256; 512 × 512
	StyleGAN2 [32]	FO	7000		256 × 256; 512 × 512; 1024 × 1024
	StyleGAN3 [30]	F	1000		256 × 256; 512 × 512; 1024 × 1024
DM	DALL-E 2 [41]	FO	3421	15.421	512 × 512; 1024 × 1024
	DALL-E MINI	O	1000		256 × 256
	Glide [39]	O	2000		256 × 256
	Latent Diffusion [42]	FO	4000		256 × 256; 512 × 512
	Stable Diffusion	FO	5000		256 × 256; 512 × 512
Nature	**Sources**	**Type**	**# Images**	**Total**	**Different Sizes**
REAL	CelebA [36]	F	5135	19341	178 × 218
	FFHQ [31]	F	4981		1024 × 1024
	Others: [33][10]	O	9225		256 × 256; 512 × 512; 1024 × 1024

The image sizes vary considerably, ranging from 216x216 pixels up to 1024x1024 pixels, thus offering a wide spectrum of resolutions for analysis. Images with different semantics (Faces (F), Other (O)) were considered. In particular, the category 'Other' includes images with semantics other than faces, such as cars, statues, paintings, etc. For each generative architecture, special attention was paid to the internal balancing of the corresponding subset of images. This balancing was pursued in terms of both semantic content and size in order to minimise potential bias and ensure a fair representation of the different types of visual input. Fig. 1 (a) show some examples of used images. All images are in PNG format.

Initially, the dataset was divided into three parts: a first 40% was used for training and validation of the BMs (refer to Section 4.1); another 40% was used for training and validation of the complete models (refer to Section 4.2); finally the remaining 20% was used as testing dataset for both phases.

4 Proposed Method

The model proposed in this paper consists of exploiting three CNN backbones as feature extractors, which are then concatenated and processed to solve the classification task. The key idea of the model lies in the training of the three backbones, each of which is trained using a specially unbalanced dataset of images (as detailed below). The purpose of this procedure is to force each backbone to focus on finding the discriminative features, left by each type of generative model during the generation phase, contained in the images belonging to a specific class (REAL, GAN-generated, DM-generated). We give the name of 'Base Model' (BM) to backbones trained on a highly unbalanced dataset and later used as feature pullers in the complete model. Figure 1 shows the entire pipeline of the proposed method.

4.1 Training of Base Models

As mentioned above, each of the three BMs was trained using a subset of the training data set. In particular, from the original image are extracted three subsets that are somewhat unbalanced with respect to each of the classes following a 90:10 ratio. In this training phase, a pre-trained Convolutional Neural Network (CNN) standard is adapted by performing a binary classification between the predominant class and the one named 'others', composed of some images taken randomly by the other two remaining classes. Figure 1 (a) summarizes the overall process. Once the training is completed, the three BMs are first frozen, the last linear layer (delegated to the binary classification) removed so that the characteristics maps of the last convolution layer are returned as output. Our hypothesis, verified during the test phase, is that, following this training procedure, the backbones focus on the search for the main characteristics of the predominant class in order to be able to recognize their presence/absence, during inference. In conducting the experiments, the following CNNs were used as

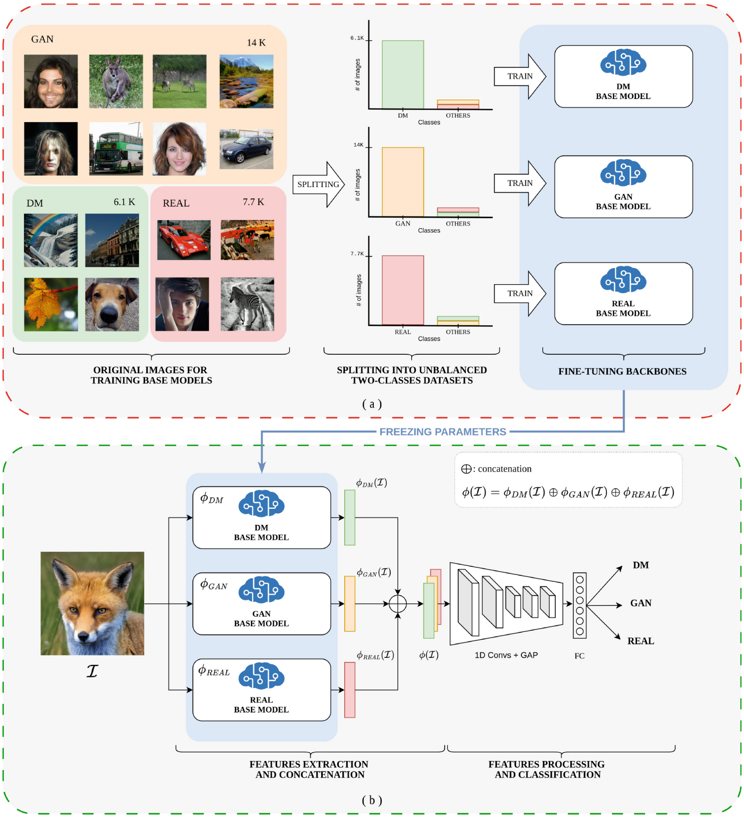


Fig. 1. Entire pipeline of the proposed method. (a) shows the process of dividing the training dataset into three unbalanced subsets, each with respect to a specific class (DM, GAN, REAL) used for training a specific BM. (b) illustrates the architecture of the final model, which takes the three BMs ϕ_c trained in the previous phase with frozen weights, and uses them to extract the features from a digital image $\phi_c(\mathcal{I})$, where $c \in \mathcal{C} = \{DM, GAN, REAL\}$. These are then concatenated in channel dimension $\phi(\mathcal{I}) = \phi_{DM}(\mathcal{I}) \oplus \phi_{GAN}(\mathcal{I}) \oplus \phi_{REAL}(\mathcal{I})$ and processed to solve the classification task.

backbone: DenseNet 121, DenseNet 161, DenseNet 169, DenseNet 201 [26], EfficientNet b0, EfficientNet b4 [47], ResNet 18, ResNet 34, ResNet 50, ResNet 101, ResNet 152 [23], ResNeXt 101 [53], ViT b16, ViT b32 [13]. All backbones have been pretrained on the Imagenet [12] dataset. All experiments were conducted on GPU NVIDIA RTX A6000 with an average training time of 90 minutes for each backbone and 210 minutes for the complete model. The parameters of each model were selected by choosing those that obtained the minimum loss value during model validation.

Table 2 shows the accuracy, recall, precision, and F1 score values obtained by evaluating all backbones on testing images. From the results we can observe how this training led to maximizing the recall value, this indicates that the classifi-

Table 2. Percentage values of the metrics Accuracy (Acc), Recall (Rec), Precision (Pre), and F1 Score (F1) obtained by testing the BM to the binary classification between the predominant class and the class 'others'.

Backbone	DM Base Model				GAN Base Model				REAL Base Model			
	Acc	**Rec**	**Pre**	**F1**	**Acc**	**Rec**	**Pre**	**F1**	**Acc**	**Rec**	**Pre**	**F1**
DenseNet 121	76.34	99.00	47.60	64.29	92.34	99.25	87.64	93.08	73.17	99.08	49.73	66.22
DenseNet 161	83.96	98.64	57.40	72.57	94.62	99.45	91.03	95.05	77.39	99.21	54.04	69.97
DenseNet 169	78.83	99.16	50.40	66.83	93.86	99.32	89.91	94.38	71.20	99.29	47.95	64.67
DenseNet 201	79.11	98.90	50.75	67.08	92.61	99.37	87.96	93.32	72.97	99.29	49.54	66.10
EfficientNet b0	85.74	97.09	60.50	74.55	88.77	98.81	82.87	90.14	77.91	97.56	54.70	70.10
EfficientNet b4	78.00	97.47	49.43	65.59	87.14	98.74	80.78	88.86	74.31	97.27	50.84	66.78
ResNet 18	76.64	98.03	47.90	64.36	84.28	99.06	77.16	86.75	65.14	99.03	43.17	60.13
ResNet 34	77.29	98.12	48.62	65.02	83.33	99.40	75.93	86.10	66.75	98.82	44.33	61.21
ResNet 50	78.14	98.61	49.59	66.00	90.34	99.14	84.82	91.42	70.79	99.13	47.59	64.31
ResNet 101	77.20	99.00	48.53	65.13	90.85	98.87	85.71	91.82	69.10	99.08	46.17	62.99
ResNet 152	76.48	99.00	47.75	64.43	93.42	99.32	89.23	94.00	70.59	98.92	47.41	64.10
ResNeXt 101	75.38	98.58	46.60	63.28	93.82	98.40	90.53	94.30	66.26	98.66	43.96	60.82
ViT b16	76.53	98.58	47.80	64.38	83.58	99.69	76.11	86.32	68.45	99.24	45.67	62.55
ViT b32	74.05	96.92	45.20	61.65	80.83	98.99	73.39	84.29	60.44	99.61	40.13	57.21

cation model is able to correctly identify all the positive examples of the interest class (the unbalanced one). In other words, the model tends to minimize false negatives; that is, there are no cases where the model wrongly classified a positive example as negative. This confirms our initial hypothesis that, following the training procedure described above, BMs are able to capture the discriminative features of each generating architecture.

4.2 Overall architecture

The final model uses the three BMs trained as described in Section 4.1 as feature extractors, at this stage they will no longer be trained as the weights have been frozen. Each BM receives the same digital image as input and is tasked with identifying and extracting the discriminative features of each class. These are then concatenated to obtain a three-channel tensor, which is then processed through a custom CNN, consisting of a sequence of 5 convolutions 1D with respectively kernel size of 7, 5, 3, 3, 3, all with padding 1 and stride 1; This was followed by a Global Average Pooling operation and a three-node linear output classifier. Figure 1 (b) presents both the entire pipeline and a graphical representation of the model.

For the training phase of the complete models we used the Cross Entropy Loss weighed with respect to the frequency of each class in the dataset (Equation 1). This choice was necessary to avoid that the models were too influenced by the imbalance present in the dataset of used images.

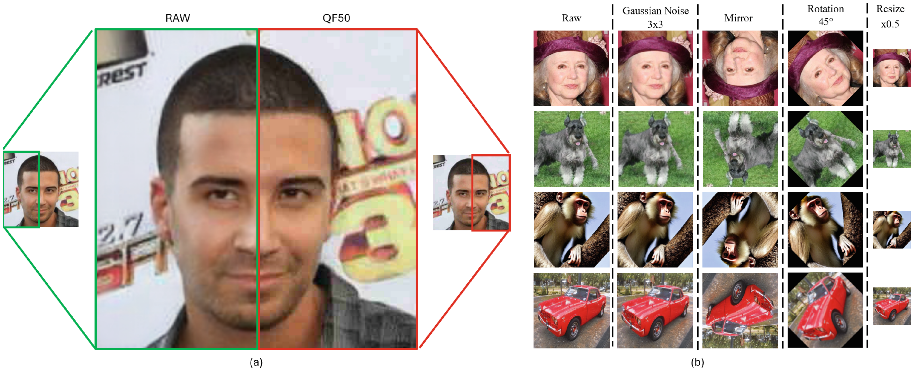


Fig. 2. Sample of images subjected to robustness attacks. (a) shows the loss of detail and quality due to JPEG compression, (b) shows various robustness attacks considered in the paper.

$$\text{Weighted Cross Entropy Loss} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{c\in\mathcal{C}} w_c y_{i,c}\log(\hat{y}_{i,c}) \tag{1}$$

where N is the number of samples, $\mathcal{C} = \{\text{GAN}, \text{DM}, \text{REAL}\}$ is the set of classes, $y_{i,c}$ is the ground truth label for sample i and c, $\hat{y}_{i,c}$ is the predicted probability for sample i and class c, and w_c is the weight for class c. In particular:

$$w_{\text{c}} = \frac{1}{\#images\ of\ class\ c} \qquad \forall c \in \mathcal{C}.$$

5 Experimental results

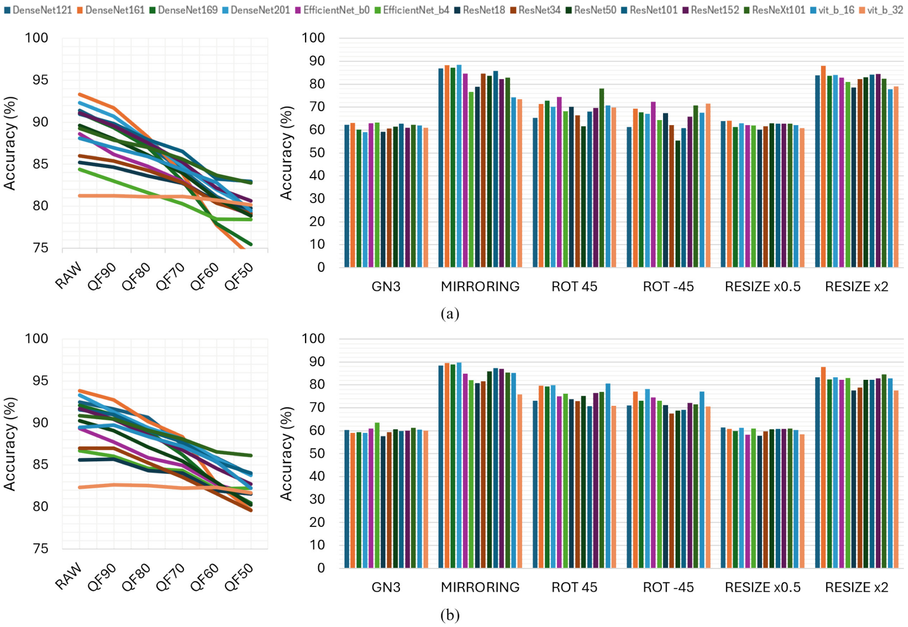
Two types of experiments were conducted: inference and robustness tests to assess the effectiveness and robustness of the classification models, and comparison with the state-of-the-art in the generalization test.

5.1 Inference and robustness tests

In this first testing phase, we tested the proposed architecture by varying the backbone of the BM in order to choose the best model. For this testing phase, 20% of the images of the original dataset (Sec. 3) were used. Furthermore, in order to make the accuracy metric more meaningful, both validation and testing datasets were balanced so as to have the same number of images for each class (DM-generated, GAN-generated, REAL).

Initially, the models were tested using only raw images. Subsequently, various attacks were applied: JPEG compression with decreasing Quality Factors (QF90 to QF50), Gaussian noise with kernel size 3 × 3 (GN3), mirroring, rotations at +45 and -45 degrees, and 2× and 0.5× factor resizing. Fig. 2 shows an example of images under some of these attacks.

In Fig. 3 (a) the performance of both tests in the three-class classification is shown. From the results obtained, it can be seen that, regardless of the backbone used in the BM, in general this approach succeeds in achieving accuracy values in excess of 85%. In particular, the use of a model belonging to the DenseNet family as a backbone gives a boost to the overall performance of the models.



(a)

(b)

Fig. 3. Performance obtained by the proposed model using different backbones in inference and robustness tests. (a) shows the metrics in three-class classification (REAL vs GAN vs DM), (b) converts these results to binary classification (REAL vs Deepfake). On the left side of each row, line graphs show how accuracy changes as image quality decreases (RAW to QF50). On the right-hand side, bar graphs compare the performance of models subjected to various robustness attacks.

To better understand the model's ability to distinguish between real and AI-generated images (from GAN or DM), we recalculated the previous performance values in the binary classification: the calculation was performed considering the predicted classes GAN and DM as deepfakes and keeping the predictions of the REAL class unchanged, then the metrics were recalculated. Fig. 3 (b) shows the metrics obtained from the recalculation.

Comparing Fig. 3 (a) and Fig. 3 (b), it's evident that the performance increased in terms of accuracy for the binary classification task. This improvement is observed both in the inference test on raw images and, notably, across the various robustness tests.

From the results obtained, DenseNet 161 represents the backbone of the Base Model as it leads to the best classification results and demonstrates good robustness at all attacks: despite the fact that the model was trained using only raw images, the accuracy values tend not to decrease drastically even after robustness attacks.

5.2 Comparison with S.O.T.A. in generalization

In this section, we examine the generalization capacity of our approach. The selected final model uses DenseNet 161 as the backbone of the BM, chosen for its excellent performance found in the tests described in Section 5.1.

Initially, we conducted an analysis of the baselines: the models used as the backbone of the BMs were trained in the same conditions of our method and subsequently evaluated in terms of generalization. This process allowed us to compare the effectiveness of our model with the use of standard architectures. Next, we compared our model with state-of-the-art models trained on similar tasks, namely the distinction between AI-generated images and real images.

In order to assess the generalisation capability of the models, we used different test sets. These test sets were divided into two categories: images generated by generative models previously observed during the training phase, but with different semantic variations and initial conditions, factors that often complicate classification, and images generated by models not included in the training phase. In addition, we conducted further tests distinguishing between images generated exclusively by GANs technologies, images generated exclusively by DMs technologies and images generated by both technologies. We use the notation whereby we define: $\mathcal{T}_*^i$ (i stays for 'inner-set') the dataset containing images generated by models already considered in the training phase; $\mathcal{T}_*^o$ (o stays for 'outer-set') the dataset containing images generated by architectures not considered in the training phase; $\mathcal{T}_*^{i/o}$ contains images generated by both type of architectures during the training phase; $\mathcal{T}_G^*$ the dataset containing only images generated by GANs as fakes; $\mathcal{T}_D^*$ the dataset containing only images generated by DMs as fakes; $\mathcal{T}_{D/G}^*$ contains images generated by both GANs and DMs architectures. Explicitly:

- $\mathcal{T}_G^i$ contains a fake image sample of 2000 divided equally between images generated by GauGAN [40], BigGAN [4], ProGAN [29], and CycleGAN [54].
- $\mathcal{T}_G^o$ contains a fake image sample of 2000 divided equally between images generated by Generative Adversarial Transformers (GANformer) [27], Denoising DiffusionGANs [52], DiffusionGANs[51], ProjectedGANs [43], and Taming Transformers [14].
- $\mathcal{T}_G^{i/o}$ contains a fake image sample of 2000 divided equally between images generated by the same generative models of $\mathcal{T}_G^i$ and $\mathcal{T}_G^o$.

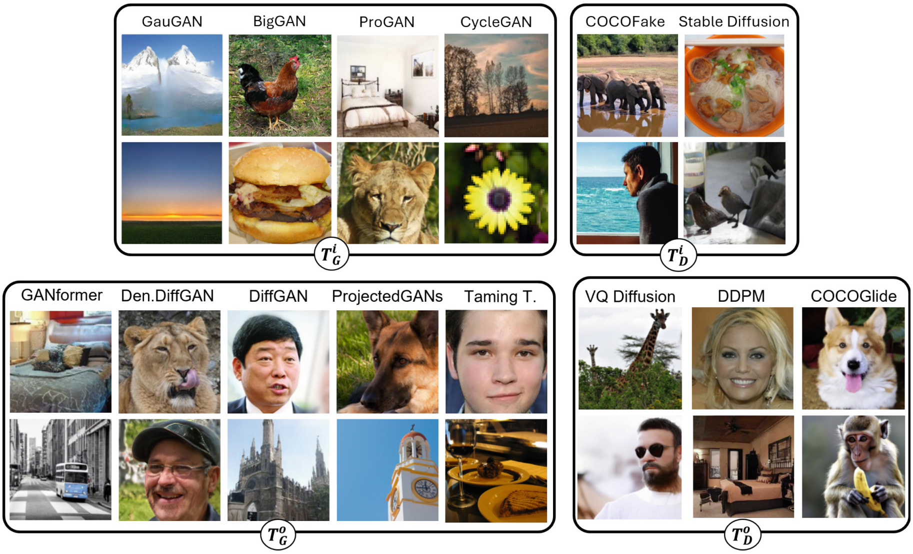


Fig. 4. Image samples generated for the generalization test. Each block also contains samples of real images taken randomly from the AFHQ [7], Imagenet [12] and COCO [34] datasets.

- $\mathcal{T}_D^i$ contains a fake image sample of 2000 divided equally between images generated by Diffusion and images taken randomly from the COCOFake dataset [8], generated by Stable Diffusion [1].
- $\mathcal{T}_D^o$ contains a fake image sample of 2000 divided equally between images generated by Vector Quantized Diffusion Model (VQ Diffusion) [18], Denoising Diffusion Probabilistic Model (DDPM) [25], and images taken randomly from the COCOGlide dataset, generated by Glide [39].
- $\mathcal{T}_D^{i/o}$ contains a fake image sample of 2000 divided equally between images generated by the same generative models of $\mathcal{T}_D^i$ and $\mathcal{T}_D^o$.
- $\mathcal{T}_{D/G}^i$ contains a fake image sample of 2000 divided equally between images generated by the same generative models of $\mathcal{T}_D^i$ and $\mathcal{T}_G^i$.
- $\mathcal{T}_{D/G}^o$ contains a fake image sample of 2000 divided equally between images generated by the same generative models of $\mathcal{T}_D^o$ and $\mathcal{T}_G^o$.
- $\mathcal{T}_{D/G}^{i/o}$ contains a fake image sample of 2000 divided equally between images generated by all the same previous generative models.

We also specify that each of the datasets listed above contains a sample of 2000 real images taken randomly in equal numbers from the datasets We also specify that each of the datasets listed above contains a sample of 2000 real images taken randomly in equal numbers from the AFHQ [7], Imagenet [12] and COCO [34] datasets. Examples of generated images are shown in Fig. 4.

[1] github.com/CompVis/stable-diffusion

Table 3. Percentage values of the accuracy obtained in generalization phase. The tests distinguished between images generated from architectures seen in the training phase, but with different initial conditions (superscript i), and images generated from architectures never seen before (superscript o), and mixed (superscript i/o). Furthermore, the tests distinguished between using only images generated by GANs (G-index), those by DMs (D-index), and mixed (G/D-index).

		$\mathcal{T}_G^i$	$\mathcal{T}_G^o$	$\mathcal{T}_G^{i/o}$	$\mathcal{T}_D^i$	$\mathcal{T}_D^o$	$\mathcal{T}_D^{i/o}$	$\mathcal{T}_{G/D}^i$	$\mathcal{T}_{G/D}^o$	$\mathcal{T}_{G/D}^{i/o}$
Baselines	DenseNet 121	56.57	**74.02**	66.96	72.07	48.20	58.68	60.58	64.23	63.13
	DenseNet 161	53.56	73.97	66.21	69.19	48.12	56.93	58.55	64.51	61.98
	DenseNet 169	52.03	66.73	61.10	65.25	43.98	52.61	56.80	57.93	57.51
	DenseNet 201	55.93	70.03	64.14	67.39	48.70	56.10	60.71	62.35	62.72
	EfficientNet b0	49.58	71.81	62.99	69.22	45.12	55.19	55.78	61.67	59.85
	EfficientNet b4	50.37	68.64	61.36	69.42	46.60	55.92	56.75	60.98	60.00
	ResNet 18	63.77	68.89	66.65	66.23	55.03	58.30	63.70	62.30	62.82
	ResNet 34	53.87	70.03	63.25	65.48	48.55	54.76	57.31	61.84	60.98
	ResNet 50	59.58	73.13	67.95	67.89	53.38	58.50	62.10	65.01	63.90
	ResNet 101	60.35	68.08	65.40	72.12	56.45	59.68	64.21	63.62	63.20
	ResNet 152	53.94	68.61	61.84	63.90	50.15	55.27	55.27	61.51	60.00
	ResNeXt 101	54.35	67.42	62.86	**74.18**	50.23	59.57	61.19	61.06	61.87
	ViT b16	65.81	73.31	69.46	68.59	52.69	58.30	**66.62**	64.53	62.72
	ViT b32	54.07	61.91	58.87	60.34	41.87	47.92	56.22	54.25	57.38
SOTA	Gandhi2020 [16]	52.30	50.79	51.71	49.91	50.86	50.34	51.54	50.57	51.06
	Wang2020 [50]	62.41	53.18	57.87	50.13	50.93	50.44	58.26	52.14	54.86
	Arshed2024 [1]	47.46	47.65	48.54	52.69	50.00	51.04	49.89	48.94	52.20
	Guarnera2024 [20]	55.00	55.63	56.23	54.11	45.98	49.97	56.07	52.21	57.17
	Our	**65.84**	72.47	**69.89**	68.09	**60.82**	**59.96**	66.06	**65.02**	**64.39**

Table 3 shows the percentage values of the accuracies obtained by the various models in the different contexts $\mathcal{T}$. When reading the results, it is important to consider that all images in the test sets are compressed in JPEG format, which, taking into account that our model was trained using only raw images, may have lowered its performance as demonstrated in Section 5.1. The state-of-the-art approaches used for comparison are [1,16,20,50]. This choice is due to the fact that almost all these methods were trained using generative architectures considered in our experiments. Wang et al. [50] and Gandhi et al. [16] used only images generated by GAN models and represent some of the best approaches in literature able to solve well the deepfake detection task (in the specific domain of GAN generated images). Despite this, experimental results reported in Table 3 show that these approaches are able to achieve similar classification results compared to methods trained considering images generated by also DM engines. However, these results show little ability to generalize. Our approach is able to generalize better, outperforming such state-of-the-art methods with classification accuracy over 10%, in any context. Arshed et al. [1] and Guarnera et al. [20] used one specific architecture to extract features for images generated by GAN and DM engines. The main limitation compared to our approach regards the strategy for feature extraction, since we used three specific models to better extract the most discriminative characteristics of the input data for each involved image category (GAN-generated, DM-generated, REAL).

In summary, from the obtained results (Table 3), our approach succeeds on average in generalizing better in most of the performed tests. Although baselines perform well in generalization when the dataset is composed of deepfake images generated by a single technology, they encounter difficulties when the dataset contains images from multiple generating architectures, both seen and unseen (column $\mathcal{T}^{i/o}_{G/D}$). Then, the proposed model outperforms all other state-of-the-art methods, confirming the good generalization ability in different contexts.

6 Conclusion and Future Works

The challenge of generalization emerges as a major obstacle in the context of deepfake detection. The ability to accurately distinguish between AI-generated and real images is crucial to monitor the ongoing development of generative models. In this article we proposed a new approach that can ensure robustness to JPEG attacks, typically used by social networks, and other various attacks, and contributed to a small step forward in solving the problem of generalization of detectors. The use of three different blocks specialized in the extraction of discriminative features of a specific images category (GAN-generated, DM-generated, real) allows our approach to develop a deeper understanding of intrinsic characteristic between real and synthetic images. This approach aims to provide a solid basis for the accurate identification of images even in the presence of variations and complexity introduced by different image generation techniques. This is the starting point for our future research: we want to strengthen the capabilities of the three discriminative feature extractors, analyze their outputs spatially and model new high-performance and structure-independent architectures.

Acknowledgements. Orazio Pontorno is a PhD candidate enrolled in the National PhD in Artificial Intelligence, XXXIX cycle, organized by Università Campus Bio-Medico di Roma. Acknowledge financial support from: PNRR MUR project PE0000013-FAIR. This research is supported by Azione IV.4 - “Dottorati e contratti di ricerca su tematiche dell’innovazione” del nuovo Asse IV del PON Ricerca e Innovazione 2014-2020 “Istruzione e ricerca per il recupero - REACT-EU”- CUP: E65F21002580005.

References

1. Arshed, M.A., Mumtaz, S., Ibrahim, M., Dewi, C., Tanveer, M., Ahmed, S.: Multiclass AI-Generated Deepfake Face Detection Using Patch-Wise Deep Learning Model. Computers **13**(1), 31 (2024)
2. Asnani, V., Yin, X., Hassner, T., Liu, X.: Reverse Engineering of Generative Models: Inferring Model Hyperparameters from Generated Images. IEEE Transactions on Pattern Analysis and Machine Intelligence (2023)

3. Bergmann, S., Moussa, D., Brand, F., Kaup, A., Riess, C.: Forensic analysis of AI-compression traces in spatial and frequency domain. Pattern Recognition Letters (2024)
4. Brock, A., Donahue, J., Simonyan, K.: Large Scale GAN Training for High Fidelity Natural Image Synthesis. In: International Conference on Learning Representations (2018)
5. Cho, W., Choi, S., Park, D.K., Shin, I., Choo, J.: Image-To-Image Translation via Group-Wise Deep Whitening-and-Coloring Transformation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10639–10647 (2019)
6. Choi, Y., Choi, M., Kim, M., Ha, J.W., Kim, S., Choo, J.: StarGAN: Unified Generative Adversarial Networks for Multi-Domain Image-to-Image Translation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 8789–8797 (2018)
7. Choi, Y., Uh, Y., Yoo, J., Ha, J.W.: StarGAN v2: Diverse Image Synthesis for Multiple Domains. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8188–8197 (2020)
8. Cocchi, F., Baraldi, L., Poppi, S., Cornia, M., Cucchiara, R.: Unveiling the Impact of Image Transformations on Deepfake Detection: An Experimental Analysis. In: International Conference on Image Analysis and Processing. pp. 345–356. Springer (2023)
9. Concas, S., Perelli, G., Marcialis, G.L., Puglisi, G.: Tensor-Based Deepfake Detection In Scaled And Compressed Images. In: 2022 IEEE International Conference on Image Processing (ICIP). pp. 3121–3125. IEEE (2022)
10. Corvi, R., Cozzolino, D., Zingarini, G., Poggi, G., Nagano, K., Verdoliva, L.: On the Detection of Synthetic Images Generated by Diffusion Models. In: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 1–5. IEEE (2023)
11. Corvi, R., Cozzolino, D., Poggi, G., Nagano, K., Verdoliva, L.: Intriguing Properties of Synthetic Images: from Generative Adversarial Networks to Diffusion Models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 973–982 (2023)
12. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition. pp. 248–255 (2009). https://doi.org/10.1109/CVPR.2009.5206848
13. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In: International Conference on Learning Representations (2020)
14. Esser, P., Rombach, R., Ommer, B.: Taming Transformers for High-Resolution Image Synthesis. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12873–12883 (2021)
15. Frank, J., Eisenhofer, T., Schönherr, L., Fischer, A., Kolossa, D., Holz, T.: Leveraging Frequency Analysis for Deep Fake Image Recognition. In: Proceedings of the 37th International Conference on Machine Learning, ICML. pp. 3247–3258. PMLR (2020)
16. Gandhi, A., Jain, S.: Adversarial Perturbations Fool Deepfake Detectors. In: 2020 International Joint Conference on Neural Networks (IJCNN). pp. 1–8. IEEE (2020)

17. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative Adversarial Nets. Advances in Neural Information Processing Systems **27** (2014)
18. Gu, S., Chen, D., Bao, J., Wen, F., Zhang, B., Chen, D., Yuan, L., Guo, B.: Vector Quantized Diffusion Model for Text-to-Image Synthesis. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10696–10706 (2022)
19. Guarnera, L., Giudice, O., Battiato, S.: Fighting Deepfake by Exposing the Convolutional Traces on Images. IEEE Access **8**, 165085–165098 (2020)
20. Guarnera, L., Giudice, O., Battiato, S.: Mastering Deepfake Detection: A Cutting-Edge Approach to Distinguish GAN and Diffusion-Model Images. ACM Transactions on Multimedia Computing, Communications and Applications (2024). 10.1145/3652027
21. Guarnera, L., Giudice, O., Nastasi, C., Battiato, S.: Preliminary Forensics Analysis of Deepfake Images. In: 2020 AEIT International Annual Conference (AEIT). pp. 1–6. IEEE (2020). 10.23919/AEIT50178.2020.9241108
22. Guarnera, L., Giudice, O., Nießner, M., Battiato, S.: On the Exploitation of Deepfake Model Recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 61–70 (2022)
23. He, K., Zhang, X., Ren, S., Sun, J.: Deep Residual Learning for Image Recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 770–778 (2016)
24. He, Z., Zuo, W., Kan, M., Shan, S., Chen, X.: AttGAN: Facial Attribute Editing by Only Changing What You Want. IEEE Trans. Image Process. **11**, 5464–5478 (2019)
25. Ho, J., Jain, A., Abbeel, P.: Denoising Diffusion Probabilistic Models. Adv. Neural. Inf. Process. Syst. **33**, 6840–6851 (2020)
26. Huang, G., Liu, Z., Van Der Maaten, L., Weinberger, K.Q.: Densely Connected Convolutional Networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 4700–4708 (2017)
27. Hudson, D.A., Zitnick, L.: Generative Adversarial Transformers. In: International Conference on Machine Learning. pp. 4487–4499. PMLR (2021)
28. Karras, T., Aila, T., Laine, S., Lehtinen, J.: Progressive Growing of GANs for Improved Quality, Stability, and Variation. In: International Conference on Learning Representations (ICLR) 2018 (2018)
29. Karras, T., Aila, T., Laine, S., Lehtinen, J.: Progressive Growing of GANs for Improved Quality, Stability, and Variation. In: International Conference on Learning Representations (2018)
30. Karras, T., Aittala, M., Laine, S., Härkönen, E., Hellsten, J., Lehtinen, J., Aila, T.: Alias-Free Generative Adversarial Networks. Adv. Neural. Inf. Process. Syst. **34**, 852–863 (2021)
31. Karras, T., Laine, S., Aila, T.: A Style-Based Generator Architecture for Generative Adversarial Networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4401–4410 (2019)
32. Karras, T., Laine, S., Aittala, M., Hellsten, J., Lehtinen, J., Aila, T.: Analyzing and Improving the Image Quality of StyleGAN. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8110–8119 (2020)
33. Leotta, R., Giudice, O., Guarnera, L., Battiato, S.: Not with My Name! Inferring Artists' Names of Input Strings Employed by Diffusion Models. In: International Conference on Image Analysis and Processing. pp. 364–375. Springer (2023)

34. Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft COCO: Common Objects in Context. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) ECCV 2014. LNCS, vol. 8693, pp. 740–755. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-10602-1_48
35. Liu, H., Li, X., Zhou, W., Chen, Y., He, Y., Xue, H., Zhang, W., Yu, N.: Spatial-Phase Shallow Learning: Rethinking Face Forgery Detection in Frequency Domain. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 772–781 (2021)
36. Liu, Z., Luo, P., Wang, X., Tang, X.: Deep Learning Face Attributes in the Wild. In: Proceedings of International Conference on Computer Vision (ICCV) (December 2015)
37. Marra, F., Gragnaniello, D., Verdoliva, L., Poggi, G.: Do GANs Leave Artificial Fingerprints? 2019 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR) pp. 506–511 (2019)
38. McCloskey, S., Albright, M.: Detecting GAN-Generated Imagery Using Saturation Cues. In: 2019 IEEE International Conference on Image Processing (ICIP). pp. 4584–4588. IEEE (2019)
39. Nichol, A.Q., Dhariwal, P., Ramesh, A., Shyam, P., Mishkin, P., Mcgrew, B., Sutskever, I., Chen, M.: GLIDE: Towards Photorealistic Image Generation and Editing with Text-Guided Diffusion Models. In: International Conference on Machine Learning. pp. 16784–16804. PMLR (2022)
40. Park, T., Liu, M.Y., Wang, T.C., Zhu, J.Y.: GauGAN: Semantic Image Synthesis with Spatially Adaptive Normalization. In: ACM SIGGRAPH 2019 Real-Time Live! pp. 1–1 (2019)
41. Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., Chen, M.: Hierarchical Text-Conditional Image Generation with CLIP Latents. arXiv preprint:2204.06125 **1**(2), 3 (2022)
42. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-Resolution Image Synthesis with Latent Diffusion Models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10684–10695 (2022)
43. Sauer, A., Chitta, K., Müller, J., Geiger, A.: Projected GANs Converge Faster. Adv. Neural. Inf. Process. Syst. **34**, 17480–17492 (2021)
44. Sha, Z., Li, Z., Yu, N., Zhang, Y.: De-fake: Detection and Attribution of Fake Images Generated by Text-to-Image Generation Models. In: Proceedings of the 2023 ACM SIGSAC Conference on Computer and Communications Security. pp. 3418–3432 (2023)
45. Shan, S., Cryan, J., Wenger, E., Zheng, H., Hanocka, R., Zhao, B.Y.: Glaze: Protecting Artists from Style Mimicry by {Text-to-Image} Models. In: 32nd USENIX Security Symposium (USENIX Security 23). pp. 2187–2204 (2023)
46. Sohl-Dickstein, J., Weiss, E., Maheswaranathan, N., Ganguli, S.: Deep unsupervised Learning Using Nonequilibrium Thermodynamics. In: International Conference on Machine Learning. pp. 2256–2265. PMLR (2015)
47. Tan, M., Le, Q.: Efficientnet: Rethinking Model Scaling for Convolutional Neural Networks. In: International Conference on Machine Learning. pp. 6105–6114. PMLR (2019)
48. Vyas, N., Kakade, S.M., Barak, B.: On Provable Copyright Protection for Generative Models. In: International Conference on Machine Learning. pp. 35277–35299. PMLR (2023)
49. Wang, R., Juefei-Xu, F., Ma, L., Xie, X., Huang, Y., Wang, J., Liu, Y.: FakeSpotter: a Simple Yet Robust Baseline for Spotting AI-Synthesized Fake Faces. In:

Proceedings of the Twenty-Ninth International Conference on International Joint Conferences on Artificial Intelligence. pp. 3444–3451 (2021)

50. Wang, S.Y., Wang, O., Zhang, R., Owens, A., Efros, A.A.: CNN-Generated Images are Surprisingly Easy to Spot... for Now. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8695–8704 (2020)
51. Wang, Z., Zheng, H., He, P., Chen, W., Zhou, M.: Diffusion-GAN: Training GANs with Diffusion. arXiv preprint arXiv:2206.02262 (2022)
52. Xiao, Z., Kreis, K., Vahdat, A.: Tackling the Generative Learning Trilemma with Denoising Diffusion GANs. arXiv preprint arXiv:2112.07804 (2021)
53. Xie, S., Girshick, R., Dollár, P., Tu, Z., He, K.: Aggregated Residual Transformations for Deep Neural Networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 1492–1500 (2017)
54. Zhu, J.Y., Park, T., Isola, P., Efros, A.A.: Unpaired Image-To-Image Translation Using Cycle-Consistent Adversarial Networks. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 2223–2232 (2017)

Generalized Image-Based Deepfake Detection Through Foundation Model Adaptation

Tai-Ming Huang[1], Yue-Hua Han[1], Ernie Chu[1], Shu-Tzu Lo[2],
Kai-Lung Hua[2,3], and Jun-Cheng Chen[1](✉)

[1] Research Center for Information Technology Innovation, Academia Sinica,
Taipei, Taiwan
pullpull@citi.sinica.edu.tw
[2] National Taiwan University of Science and Technology, Taipei, Taiwan
[3] Microsoft, Taipei, Taiwan

Abstract. With the rapid advancement of image synthesis and manipulation techniques from Generative Adversarial Networks (GANs) to Diffusion Models (DMs), the generated images, often referred to as Deepfakes, have been indistinguishable from genuine images by human and thus raised the public concerns about potential risks of malicious exploitation such as dissemination of misinformation. However, it remains an open and challenging task to detect Deepfakes, especially to generalize to novel and unseen generation methods. To address this issue, we propose a novel generalized Deepfake detector for diverse AI-generated images. Our proposed detector, a side-network-based adapter, leverages the rich prior encoded in the multi-layer features of the image encoder from Contrastive Language Image Pre-training (CLIP) for effective feature aggregation and detection. In addition, we also introduce the novel **Di**versely **GEN**erated image dataset (**DiGEN**), which encompasses the collected real images and the synthetic ones generated from versatile GANs to the latest DMs, to facilitate better model training and evaluation. The dataset well complements the existing ones and contains sixteen different generative models in total over three distinct scenarios. Through extensive experiments, the results demonstrate that our approach effectively generalizes to unseen Deepfakes, significantly surpassing previous state-of-the-art methods. Our code and dataset are available at https://github.com/aiiu-lab/AdaptCLIP.

Keywords: Deepfake Dataset · Deepfake Detection · Foundation Model Adaptation · Diffusion Model

1 Introduction

Image synthesis has been a popular research topic in computer vision since the introduction of Generative Adversarial Networks (GAN) [9], which can generate photo-realistic images indistinguishable from real-world photos by human.

A. Antonacopoulos et al. (Eds.): ICPR 2024, LNCS 15321, pp. 194–210, 2025.
https://doi.org/10.1007/978-3-031-78305-0_13

Recently, Diffusion Models (DM) [10] have further advanced the image generation quality and diversity over GAN-based methods. In addition, many open sourced text-to-image (text2image) models (e.g., Stable Diffusion [28]) and commercial online services (e.g., Midjourney[1]) based on DMs showcase the remarkable generation results. However, these tools also raise the public concerns about the potential privacy and copyright issues. To address the concerns, synthetic image detection, also known as generalized image-based Deepfake detection, has been proposed and researched for years. Nevertheless, it is still an open and challenging problem to develop an effective detector generalizing well on newly developed unseen Deepfakes. Recently, although there exists some works [5,23] proposing to adapt the backbone of the foundation model (i.e., Contrastive Language Image Pre-training (CLIP) [26] which has been trained on 400 million image-text pairs and exhibits strong zero-shot and few-shot performance across various task.) for the generalized image-based Deepfake detection task, their approaches often underutilize the rich information of CLIP, leading to suboptimal performance. On the other hand, the task also requires a new dataset for a more comprehensive performance evaluation of generalized image-based Deepfake detection upon the synthetic images generated by more recent GANs and DMs.

To handle these issues, we first introduce a new **Di**versely **GEN**erated image dataset (**DiGEN**) to cover more scenes and generative models. The dataset contains the real images and the synthetic images generated by the unconditional models pre-trained upon the LSUN-Bedroom [38] and FFHQ [13] datasets and the conditional text-to-image ones upon the MSCOCO [15] dataset, respectively, involving sixteen recent generative models in total (i.e., seven GANs and nine DMs.). The proposed dataset not only complements the existing ones with the latest publicly available DMs but also provides a unified evaluation protocol.

Furthermore, we propose a novel generalized image-based Deepfake detector. This detector utilizes the multi-layer features of CLIP through our proposed side-network-based adaptation method, enabling superior adaptation of the foundational model for better generalized image-based Deepfake detection than previous approaches. With extensive experiments, the evaluation results demonstrate the proposed method not only generalizes well to various unseen Deepfakes but also outperforms other previous state-of-the-art methods by a large margin. To sum up, the main contributions of this paper are summarized as follows:

- We introduce the **DiGEN** dataset and a unified evaluation protocol, where the dataset encompasses the generated images from sixteen recent generative methods under three different scenarios.
- We propose a novel and effective side-network-based adapter of CLIP for generalized image-based Deepfake detection. The proposed model exhibits strong generalization to unseen Deepfakes, outperforming other state-of-the-art methods by a large margin.

[1] https://www.midjourney.com

2 Related works

Generalized Deepfake Image Dataset. In recent years, an increasing number of synthetic image generation methods have been introduced, accompanied by the emergence of many datasets consisting of AI-generated images, such as CNNDet [35], are solely based on the images produced by GAN. Sha et al. [31] present a dataset derived from images generated by four distinct text-to-image DMs. Similarly, Bird et al. [1] present a dataset consisting of general images solely generated by Stable Diffusion 1.4[2]. Ricker et al. [27] establish a diverse dataset that includes images generated by both GANs and DMs. Wang et al. [36] present a dataset comprising two distinct scene categories and nine DM architectures. Nevertheless, with the rapid and lasting development of new generative models, generalized image-based Deepfake detection is still a challenging problem, underscoring the continuing need for a new dataset to cover more recent generative model in more diverse scenarios.

Generalized Image-based Deepfake Detection. In response to the growing security and privacy apprehensions regarding generative models, numerous studies have been conducted to tackle the challenge of identifying AI-generated images through various approaches, including image-based methods [17,36,40], frequency-based approaches [25,33], and the methods based on pre-trained models [23]. Wang et al. [35] have first shown that the detector trained upon the synthetic images generated by ProGAN [11] generalizes well to other unseen GAN-generated images, but this approach still fails to detect recent DM-generated ones. Many other works [3–5,7,39] have recently been proposed to further advance the detection performance and to highlight the structural differences between DMs and previous generative methods. Recent studies have focused on leveraging a fixed foundation model that, having been pre-trained on large-scale datasets, encodes a wealth of information within its features. This approach has shown to offer superior zero-shot or few-shot performance across various downstream tasks. For instance, Gao et al. [8] develop feature adapters to fine-tune the features from the encoders of foundational model, i.e., CLIP, tailoring them for specific applications. Similarly, Ojha et al. [23] propose UniFD, which utilizes the rich feature space of CLIP, for Deepfake detection. They investigate the encoding capabilities of CLIP's frozen image encoder, which was trained on a vast dataset of text-image pairs, and they adapt this encoder by coupling it with a linear classifier. In this work, our motivation diverges from that of the related UniFD approach. Compared to UniFD, which tends to utilize the features from the final layer of the image encoder without any training or even employs a linear layer to fine-tune the frozen image encoder, our method utilizes a side-network-based adaptation technique to fully exploit the multi-layer features of the frozen image encoder. By conducting a thorough analysis of the multi-layer features from the frozen image encoder, we were able to pinpoint the most effective features for adaptation. Our experiments demonstrate that employing this method of adaptation markedly enhances the generalization capabilities for detecting Deepfakes.

[2] https://huggingface.co/CompVis/stable-diffusion-v1-4

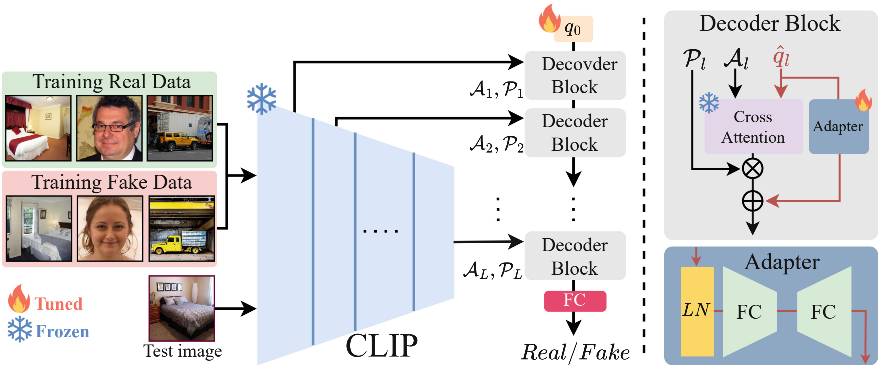


Fig. 1. Framework Overview: Our method adapts the frozen CLIP image encoder using the proposed side-network-based adapter, consisting of a series of decoder blocks, to the downstream task of generalized image-based Deepfake detection. Through this design of the multiple cascaded decoder blocks, our approach can progressively refine the query tokens to fully exploit the rich features across various layers of the frozen image encoder, ensuring better adaptation. Ultimately, the learnable query token is transmitted from the top-most to the bottom-most decoder block, culminating in a final prediction via the classification head.

3 Method

This section begins with an introduction to our DiGEN dataset, detailed in Section 3.1. Subsequently, in Section 3.2, we demonstrate the details of the proposed side-network-based adapter to fine-tune the foundation model (i.e., CLIP) for better generalized image-based Deepfake detection. An overview of our framework is illustrated in Fig. 1.

3.1 DiGEN Dataset

First, to better assess the performance of a generalized image-based Deepfake detector considering both traditional GANs and the lastest DMs, we collect a diverse dataset, DiGEN, composed of images generated by sixteen different generative models, seven GANs and nine DMs, in three different scenarios, including LSUN-Bedroom [38], FFHQ [13], and MSCOCO [15], also serve as the three subsets for the experiment. The details of our dataset and the comparisons with others are presented in Table 1 and Table 2, respectively. We also showcase qualitative results, as illustrated in Fig. 2. Furthermore, we provide a detailed description of the three scenarios as follows:

Table 1. Generation details for each image category of the proposed DiGEN dataset. We provide comprehensive documentation of the generative models and the corresponding quantities of real and fake images for each image category in our DiGEN dataset. Additionally, we report the Fréchet Inception Distance (FID) metric to evaluate the quality of the generated data.

Image Category	Generator	# of images (real/generated)	FID
LSUN-Bedroom [38] (LSUN-B)	ADM [6]	50K/50K	1.90
	DDPM [10]	50K/50K	6.36
	IDDPM [22]	50K/50K	4.24
	PNDM [18]	50K/50K	5.68
	LDM [29]	50K/50K	3.42
	ProGAN [11]	50K/50K	8.34
	StyleGAN [13]	50K/50K	2.65
	Projected-GAN [30]	50K/50K	1.52
	Diff-StyleGAN2 [37]	50K/50K	3.65
	Diff-ProjectedGAN [37]	50K/50K	1.43
	LDM [29] (text2image)	10K/10K	71.11
	SD v1.4 [28] (text2image)	10K/10K	44.37
	SD v2.0 [28] (text2image)	10K/10K	45.06
	SDXL v1.0 [24] (text2image)	10K/10K	54.99
FFHQ [13]	ADM* [2]	50K/50K	6.92
	LDM [29]	50K/50K	4.98
	Projected-GAN [30]	50K/50K	3.39
	Diff-StyleGAN2 [37]	50K/50K	2.83
	StyleGAN [13]	50K/50K	4.40
	StyleGAN v2 [14]	50K/50K	2.84
	StyleGAN v3 [12]	50K/50K	3.07
MSCOCO [15]	SD v1.4 [28] (text2image)	60K/60K	20.61

Table 2. Comparisons with other existing datasets. We compare DiGEN with existing datasets according to three major criteria: Category, Generator, Public Access, and Real/Fake Images.

Dataset	Category	Generator	Public Access	Real/Fake
CNNDet	LSUN, General	1 GANs	Yes	362K/362K
DE-FAKE	General	4 DMs	No	20K/60K
CiFAKE	General	1 DMs	No	60K/60K
Towards.	LSUN-B	5 GANs, 5 DMs	Yes	500K/500K
DIRE	LSUN-B, Face, General	1 GANs, 9 DMs	Yes	488K/134K
DiGEN	LSUN-B, Face, General	**7 GANs, 9 DMs**	**Yes**	**950K/950K**

1) LSUN-Bedroom (LSUN-B). Following the protocol of [35], we expand the dataset[3] which includes images from five GAN models (ProGAN [11], StyleGAN [13], etc.) and five DM models (DDPM [10], iDDPM [22], etc.) pre-trained on LSUN-Bedroom [38]. Specifically, we complements the prior datasets with the latest text-to-image DM frameworks (LDM [29], SD-v1[4], SD-v2[5], even the most advanced SDXL-v1.0 [24]) and the prompt "A photo of a bedroom" for synthetic image generation. For unconditionally generated images, each model produces 50K images, with 39K designated as the training set, 1K as the validation set, and 10K as the test set. For the images generated using text-to-image models, we collect 10K images each for the test set.

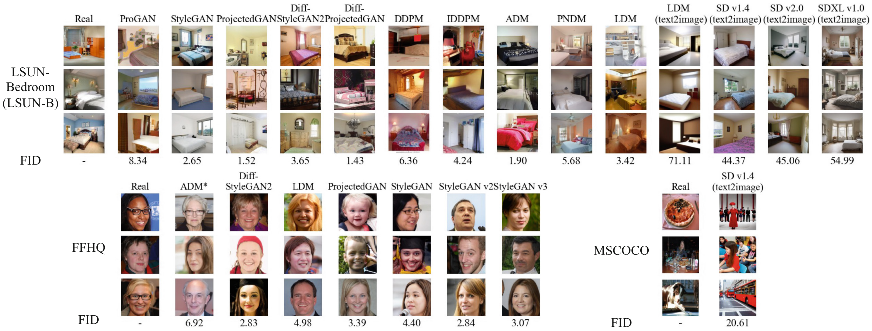


Fig. 2. Uncurated samples from the DiGEN dataset. We present a diverse selection of images randomly sampled from our DiGEN dataset, showcasing the output of various generative models across different scenarios.

2) FFHQ. DiGEN also aims to provide more diverse scenarios (e.g., face) with higher quality data than other similar ones (e.g., CelebA-HQ subset in DIRE [36]). Thus, we have specifically collected images generated by more recent and publicly available pre-trained weights on the FFHQ facial dataset. These consist of seven unconditional generation models (ADM* [2], LDM [29], Projected-GAN [30], Diff-StyleGAN2, Diff-ProjectedGAN [30], StyleGAN [13], StyleGANv2 [14], StyleGANv3 [12], and ADM* is a smaller version of ADM with 93M instead of more than 500M parameters due to the available pre-trained weights.). We have collected 50K generated images from each of these models, divided into 39K for training, 1K for validation, and 10K for testing.

3) MSCOCO. For more comprehensive and unbiased evaluations, we collect "Natural" scene images using text-to-image DM SD-v1 [28]. Instead of using "A photo of {class name}" as a prompt, we employ the complete image caption from the MSCOCO dataset to guide the generation.

[3] https://github.com/jonasricker/diffusion-model-deepfake-detection

[4] https://huggingface.co/CompVis/stable-diffusion-v1-4

[5] https://huggingface.co/stabilityai/stable-diffusion-2-1

3.2 Foundation Model Adaptation for Generalized Image-based Deepfake Detection

[6] Recent studies [16,23] highlight the powerful zero-shot capabilities of CLIP's image encoder, which are attributed to its text-image multimodal training framework. This multimodal approach endows the model with robust semantic understanding and encoding capabilities. CLIP model's image encoder, trained on an extraordinarily large dataset of 400M image-text pairs, has gained extensive exposure to the visual world, resulting in formidable visual encoding capabilities. This makes it an ideal candidate for addressing various downstream tasks. In contrast to directly employing CLIP model's image encoder feature space or using linear probing for transfer learning [23], we posit that the frozen backbone encodes valuable implicit information. To effectively leverage the feature space of the foundation model, we present our novel lightweight side-network adaptation method to facilitate the model's improved generalization in image-based Deepfake detection tasks. As depicted in Fig. 1, we reprogram CLIP model's frozen image encoder by pairing it with multiple decoder blocks. The latents encoded by the image encoder encapsulate rich and versatile features, while our multi-layer adapters capture general Deepfake cues implicit at different levels. This adaptation strategy allows us to maximize the utilization of rich features from each layer of the foundational model, thereby enhancing the generalization capability of our image-based Deepfake detection task.To implement our multi-layer adaptation strategy, we extract specific features from each layer of the CLIP model's image encoder. These features serve as the basis for our AdaptCLIP method, allowing us to fully utilize the information encoded in the frozen backbone.

Moreover, from each layer of the ViT-based image encoder, we export patch embeddings $\mathcal{P}_l \in \mathbb{R}^{P\times(H\times D)}$ and an attention attribute $\mathcal{A}_l \in \mathbb{R}^{P\times H\times D}$, representing one of the attention attribute types (i.e., query, key, and value), where $1 \leq l \leq L$ indicates the layer index, L the number of transformer layers, P the number of patches, H the number of attention heads in the multi-head self-attention mechanism, and D the feature dimension per head for each transformer layer of the image encoder.

We consider q_0 as the initial learnable query token parameters, learning from the top layer to the bottom, then projecting to a classification prediction via a linear layer. Formally, the operations of the adaptation method can be formulated as:

$$q_l = DecoderBlock(q_{l-1}, \mathcal{P}_l, \mathcal{A}_l),\ l = 1, ..., L \tag{1}$$

$$p = softmax(FC(q_L)), \tag{2}$$

where P_l and A_l are extracted from the l_{th} layer of the frozen image encoder. q_l is the progressively tuned query token and $FC(\cdot)$ is a linear layer that projects the final query token q_L to class predictions. Finally, p represents the final prediction results, indicating the probability distribution for a prediction being either *Real*

[6] https://github.com/crywang/face-forgery-detection

Table 3. Performance comparisons of our AdaptCLIP and other generated image detectors. * denotes we retrain the detection models using their official codes upon the synthetic images generated by ADM which is trained on LSUN-B. In the case of TwoStream[20], we produce the results from an unofficial github repository[9]due to the official code is unavailable. We report the evaluation results in ACC (%)/AP (%).

Detection method	Training dataset	Generation model	Testing DMs					Testing GANs					Total Avg.
			ADM	DDPM	IDDPM	PNDM	LDM	Pro-GAN	Style-GAN	Proj.-GAN	Diff-StyleGAN2	Diff-Proj.GAN	
CNNDet	LSUN	ProGAN	50.1/66.4	50.3/82.5	50.1/78.9	50.1/77.5	50.2/75.9	99.7/**100**	59.2/97.1	52.6/92.6	80.9/99.7	51.6/91.2	59.5/86.2
GANDet	LSUN	ProGAN	50.0/61.2	50.0/59.0	50.1/64.6	50.4/73.4	50.1/56.9	54.4/89.9	51.4/90.0	50.1/58.5	95.1/99.6	50.3/62.8	55.2/71.6
SBI	FF++	Multiple	49.5/49.3	50.2/50.0	50.7/50.6	49.6/49.9	50.2/50.2	50.0/50.2	49.8/50.2	49.4/49.5	50.1/50.0	50.8/50.9	50.0/50.1
TwoStream	FF++	Multiple	50.0/49.9	50.0/52.3	50.0/50.9	50.0/56.2	50.0/50.3	50.0/53.4	50.0/53.0	50.0/47.9	50.0/52.9	50.0/49.2	50.0/51.6
DIRE	LSUN-B.	ADM	50.2/63.5	50.2/64.1	50.3/64.5	50.1/56.4	50.2/61.1	50.3/63.2	50.2/58.8	50.3/62.5	50.3/66.1	50.2/61.1	50.2/62.1
UniFD	LSUN	ProGAN	53.9/81.3	65.9/95.9	62.5/91.5	74.5/97.1	66.8/94.4	97.7/99.9	60.1/92.4	81.8/98.8	83.2/98.4	76.5/98.4	72.3/94.8
FreqNet	LSUN	ProGAN	52.7/87.5	81.4/99.3	55.5/90.1	56.6/91.8	99.4/**100**	99.1/**100**	89.3/99.6	99.8/**100**	94.5/99.9	99.7/**100**	82.8/96.2
CNNDet*	LSUN-B	ADM	92.8/99.4	93.3/99.5	98.4/99.9	95.4/99.8	85.4/98.7	94.3/98.8	82.5/97.5	63.5/82.6	61.5/87.4	61.8/85.6	82.9/94.9
DIRE*	LSUN-B.	ADM	97.1/99.9	94.5/98.9	96.9/99.5	92.5/97.9	74.2/91.4	67.4/88.9	60.2/83.6	64.7/87.5	66.0/86.9	62.1/86.2	77.5/92.1
UniFD*	LSUN-B.	ADM	82.2/90.9	91.5/98.7	89.8/97.2	92.1/99.3	91.5/98.8	92.4/99.9	89.4/96.1	91.3/98.6	91.7/99.0	91.6/98.5	90.3/97.7
FreqNet*	LSUN-B.	ADM	**99.1**/**100**	**99.0100**	99.1/**100**	50.0/45.2	66.0/93.3	87.4/99.0	78.1/94.3	87.4/99.0	73.9/89.4	87.2/99.0	83.3/91.9
AdaptCLIP (Ours)	LSUN-B.	ADM	94.3/99.7	**99.0**/**100**	**99.4**/**100**	99.8/**100**	99.8/**100**	99.8/**100**	**99.3**/**100**	**99.6**/**100**	98.9/**100**	**99.6**/**100**	**98.9**/**100**
	LSUN-B.	IDDPM	87.1/97.6	**99.0**/**100**	98.7/99.9	99.0/**100**	**99.9**/99.9	99.0/**100**	97.6/99.7	97.7/98.9	98.6/**99.9**	97.4/99.7	97.4/99.6
	LSUN-B.	LDM	68.3/97.5	94.9/**99.9**	93.3/**100**	**99.9**/**100**	99.4/**100**	**100**/**100**	**97.2**/**100**	99.3/**100**	**99.2**/**100**	98.9/**100**	94.6/99.7

or $Fake$. We employ binary cross-entropy loss for adaptation training.

The $DecoderBlock(\cdot)$ is designed to aggregate and utilize the cues from the image encoder's generalizable features with a learnable query. Formally,

$$DecoderBlock(q, \mathcal{P}, \mathcal{A}) = \hat{\mathcal{M}} + Adapter(q), \tag{3}$$

$$\text{where } \hat{\mathcal{M}} = \mathcal{M} \cdot \mathcal{P}, \text{and } \mathcal{M} = CA(q, \mathcal{A}), \tag{4}$$

with the query q from the previous layer, we employ Cross-Attention ($CA(\cdot)$) on the attribute $\mathcal{A}$ to produce an affinity map $\mathcal{M}$. We apply matrix multiplication between the affinity map $\mathcal{M}$ and the patch embeddings $\mathcal{P}$ to produce the intermediate query feature map $\hat{\mathcal{M}}$. At the core of $DecoderBlock$, an $Adapter(\cdot)$, parameterized by two fully connected (FC) layers with a GELU activation in-between, transforms q into forgery-aware features, which are then summed with $\hat{\mathcal{M}}$ and passed to the next layer. We apply the layer normalization (LN) on q before feeding it to the $Adapter(\cdot)$. In summary, the aggregated query collects all the rich information from the frozen features of the l-th encoder layer and passes it to the $(l + 1)$-th blocks.

4 Experiment

In this section, we first introduce our experimental setup followed by extensive experimental results in various scenarios and ablation studies.

4.1 Experimental Setup

Implementation Details. For our base model, we choose CLIP ViT-L/14 for its superior performance. Our training details include using AdamW [19] optimizer with a batch size of 256 and a learning rate of 1e-3, along with cross-entropy loss to guide model convergence. Training spans 30 epochs, incorporating early stopping with a patience of 5 steps to preserve the best weights.

Evaluation Metrics. We report accuracy (ACC) and the average precision (AP). The threshold to compute ACC is set to 0.5.

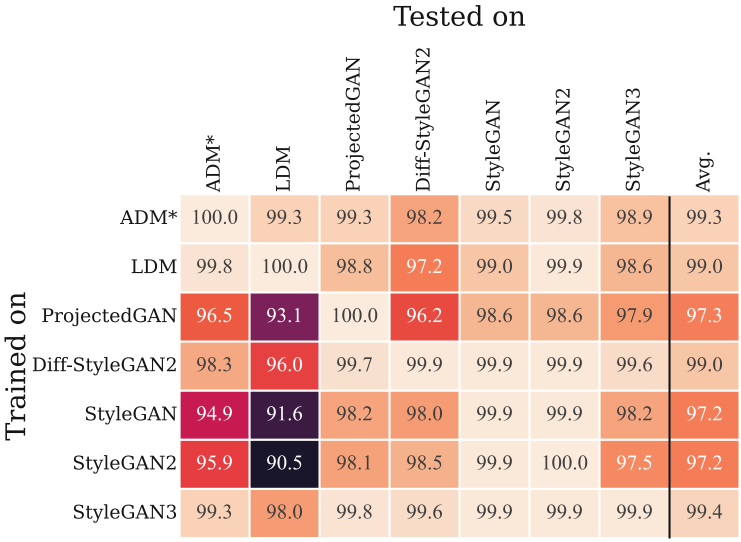
Baseline Models. We adopt the following methods for comparisons.

1. **CNNDet** [35]: Wang et al. propose a generalized image-based Deepfake detector trained using ProGAN-generated images.
2. **GANDet** [21]: GANDet employs an ensemble of multiple EfficientNet-B4 [34] models to enhance the generalization performance of Deepfake detection.
3. **SBI** [32]: SBI adopts a self-supervised learning method, simulating fake images from real images to achieve superior detection performance.
4. **TwoStream** [20]: It utilizes high-frequency features and residual-guided fusion for improved generalized Deepfake detection.
5. **DIRE** [36]: DIRE introduces a novel image representation technique that assesses the discrepancy between an input image and its reconstructed counterpart using a pre-trained diffusion model. This serves as a bridge to differentiate between the real and generated images.
6. **UniFD** [23]: UniFD employs pre-trained CLIP [26] models, utilizing its rich features alongside the nearest neighbor classifier and linear probing techniques to distinguish between real and fake.
7. **FreqNet** [33]: FreqNet focuses on learning the generalized cues in the frequency domain for Deepfake detection.

Additionally, we retrain the models of CNNDet, DIRE, UniFD, and FreqNet with their official codes and settings on the images generated by ADM which is trained on LSUN-B (i.e., we denotes it as LSUN-B-ADM subset of the DiGEN dataset.) and subsequently evaluate their performances.

4.2 Comparisons with Other State-of-the-art Detectors

Recent advancements in generative models have notably improved image generation. While prior studies detect such images well, few explore the newer models' generalization. In this study, we assess the performance of several notable detectors—CNNDet [35], GANDet [21], SBI [32], TwoStream [20], DIRE [36], UniFD [23], and FreqNet [33]—on the LSUN-B subset of the DiGEN dataset. The quantitative results can be found in Table 3. All detectors show decreased performance against novel diffusion models, indicating DM-generated images differ structurally from past GAN-generated ones, resulting in reduced generalization for unseen images. We utilize images generated by ADM [6] as our training set, subsequently retraining baseline models for the same number of steps as outlined in their official open-source implementation. * indicates our LSUN-B-ADM subset training on DiGEN using official codes. Despite encountering



	ADM*	LDM	ProjectedGAN	Diff-StyleGAN2	StyleGAN	StyleGAN2	StyleGAN3	Avg.
ADM*	100.0	99.3	99.3	98.2	99.5	99.8	98.9	99.3
LDM	99.8	100.0	98.8	97.2	99.0	99.9	98.6	99.0
ProjectedGAN	96.5	93.1	100.0	96.2	98.6	98.6	97.9	97.3
Diff-StyleGAN2	98.3	96.0	99.7	99.9	99.9	99.9	99.6	99.0
StyleGAN	94.9	91.6	98.2	98.0	99.9	99.9	98.2	97.2
StyleGAN2	95.9	90.5	98.1	98.5	99.9	100.0	97.5	97.2
StyleGAN3	99.3	98.0	99.8	99.6	99.9	99.9	99.9	99.4

Fig. 3. Detection performance for the proposed AdaptCLIP trained on the FFHQ face subset of DiGEN dataset. The rows and columns respectively represent models trained on and tested on samples from various GANs and DMs.

generalization challenges, our proposed AdaptCLIP method exhibited remarkable zero-shot generalization capabilities across nine different architectures. It achieved average ACC and average AP scores of 98.9% and 100%, respectively, even when applied to GAN-generated images. In the following two subsections, we provide more evaluation results across different domains

4.3 Cross-Facial Domain Generalization over FFHQ.

Given the growing concerns of Deepfakes generated by AI, facial privacy threats are becoming increasingly alarming. We also verify the effectiveness of the proposed method in the FFHQ scenario. Employing the facial subset mentioned in Section 3.1, we train our model on the images from seven different generative methods and conduct zero-shot evaluation on six others, with the results depicted in Fig. 3. The results show that by adapting the foundation model, we can effectively apply our method to different domains for detection. (i.e., in this setting, both training and test sets are in the FFHQ scenario.)

4.4 Cross-Content Domain Generalization.

Besides the cross-facial domain evaluation, we further design a challenging cross-content domain evaluation to train on images from an unconditional generative method (the ADM subset of the Generated-LSUN-B) and evaluate the generalization towards text-to-image generative methods (the Generated-LSUN-B subset with four GAN and DM models; the Generated-FFHQ subset with seven GAN and DM models; and the Generated-MSCOCO subset with 80 real-world categories).

Adaptation Method Effectiveness To validate the efficacy of our proposed method, we conduct a comprehensive comparative analysis against other baseline models and various adaptation techniques under the cross-domain generalization scenario: (a) CNNDet, which involves retraining from scratch as per the approach in [35], (b) UniFD, also fine-tuned following the settings as in [23], (c) EVL [16], which employs multi-layer frozen features from CLIP for enhanced video recognition capabilities, (f) LinearCLIP (ViT-L/14) utilizing a straightforward linear layer and layer normalization for the classification task. In contrast to UniFD, LinearCLIP directly harnesses the features from the final decoder block without projecting them into a 768-dimensional feature space. (d), (e), and (g) denote our AdaptCLIP using the available pre-trained CLIP image encoders in different backbone architectures, including ViT-B/16, ViT-B/32, and ViT-L/14. The results are shown in Table 4. Our AdaptCLIP methodology, which

Table 4. Evaluation of different adaptation methods. We conduct experiments with various foundation model adaption methods to evaluate the AP(%) for cross-domain performance. * denotes the retrained on the LSUN-B-ADM subset of the DiGEN dataset using the official codes.

Adaptation Method	Bedroom (Text2Image)	FFHQ	MSCOCO	Avg.
(a) CNNDet*	60.3	63.4	45.5	56.4
(b) UniFD*	86.8	88.3	86.4	87.2
(c) EVL*	88.3	90.6	75.0	84.6
(d) AdaptCLIP (ViT-B/32)	58.1	58.3	60.1	58.8
(e) AdaptCLIP (ViT-B/16)	73.0	67.7	64.2	68.3
(f) LinearCLIP (ViT-L/14)	85.1	89.5	88.5	87.7
(g) AdaptCLIP (ViT-L/14)	**89.5**	**92.8**	**93.1**	**91.8**

Table 5. Evaluation of feature attribute selection. To assess our model's effectiveness, we meticulously chose a variety of features—specifically q, k or v —from the frozen image encoder. This selection process is critical for verifying the robustness and adaptability of each attribute across different domains.

$\mathcal{A}_l$	Bedroom (Text2Image)	FFHQ	MSCOCO	Avg.
$\mathcal{A}_l = k$	**89.5**	**92.8**	**93.1**	**91.8**
$\mathcal{A}_l = q$	82.7	83.9	80.9	82.5
$\mathcal{A}_l = v$	84.1	57.7	65.7	69.2

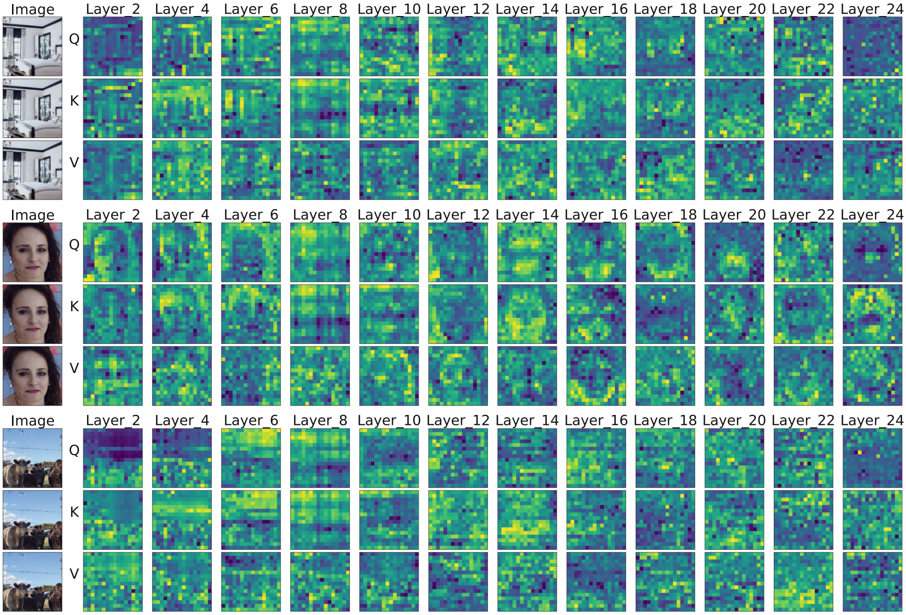


Fig. 4. Visualization of the features from the image encoder of the frozen CLIP [26]. We present comprehensive visualizations of the q, k and v features from the frozen image encoder across multiple layers, focusing on images from three distinct subsets of the DiGEN dataset: LSUN-B, FFHQ, and MSCOCO. Our goal is to meticulously examine how features from various layers and attributes influence the effectiveness of our proposed method. This analysis sheds light on the nuanced impact of different feature layers on model performance, providing valuable insights into the adaptability and robustness of our approach.

adeptly harnesses multi-layer features from foundational features, outperforms other approaches, demonstrating its superior efficacy in cross-domain generalization scenario. Moreover, our experiments reveal the method's remarkable adaptability to CLIP image encoders of various sizes, highlighting a trend where larger models tend to achieve better generalization. This finding not only showcases the versatility of AdaptCLIP but also affirms the strategic advantage of leveraging foundational model features to enhance detection capabilities.

Feature Selection from Frozen Backbone In our analysis, we commence by illustrating the diverse attributes of the frozen features at each layer within the image encoder, as depicted in Fig. 4. For clearer presentation, we showcase the attributes every two layers instead of every layer. Specifically, we meticulously extract and normalize the multi-layer frozen features of q, k, and v from the image encoder to present them. These results are then distinctly showcased for the LSUN-B, FFHQ, and MSCOCO subsets of the DiGEN dataset. Our analysis uncovers that each layer of the encoder contains a wealth of unique information: in the shallower layers, the features focus on fine-grained details, while in the deeper layers, they emphasize semantic information of the whole image.

Additionally, we find that for q, k, and v, each captures different types of information, showcasing the complex and multi-dimensional nature of the encoding process.

In Table 5, we conduct an examination of various attribute selections to determine their impact on leveraging the image features of the frozen CLIP for improving the generalizability of our model in the image-based Deepfake detection task. The findings highlight that selecting either q or k as attribute boosts the model's generalization capabilities, while v exhibits worse performance. Our assumption is based on the inherent design of q and k within the attention block to foster affinity, in contrast, opting for v leads to a notable decrease in performance since it is not intended for the purpose.

Furthermore, we have included an investigation into the number of frozen layers, as shown in Table 5. In our experiment, we select layers at regular intervals. Specifically, when $L = 12$, we select layers $0, 2, ..., 22$; when $L = 6$, we select layers $0, 4, ..., 20$, and so forth. We find that using all layers ($L = 24$, as CLIP:ViT-L/14 is our selected foundational model.) yields the best general-

Table 6. Evaluation of forzen layers selection. We investigated the impact of adapting different numbers of frozen layers on the model's generalization capability. Our experiments revealed that increasing the number of adapted layers led to improved generalization performance, thereby demonstrating the effectiveness of our method.

L	Bedroom (Text2Image)	FFHQ	MSCOCO	Avg.
$L = 24$	**89.5**	**92.8**	**93.1**	**91.8**
$L = 12$	86.3	86.8	79.0	84.0
$L = 6$	88.3	73.0	69.8	77.1

Table 7. Generalizability across proportionally scaled training datasets To confirm the effectiveness of our adaptation strategy, we conduct an evaluation focusing on how well it generalizes across domains when trained with varying quantities of training data. This assessment aims to understand the robustness and efficiency of our method under different numbers of training data.

# of Training Data	Bedroom (Text2Image)	FFHQ	MSCOCO	Avg.
10%	69.8	86.4	77.8	81.4
25%	88.1	92.0	92.1	90.7
50%	87.4	92.1	92.7	90.7
70%	86.7	92.0	91.8	90.2
100%	**89.5**	**92.8**	**93.1**	**91.8**

ization performance. Our comprehensive analysis reveals that each layer of the encoder encapsulates a wealth of unique information. Consequently, the design of AdaptCLIP effectively leverages features from all layers, significantly enhancing the generalization performance in Deepfake detection.

Data Efficiency of AdaptCLIP In our experiment, we explore how varying the number of training data influences the efficacy of our method. The findings are detailed in Table 7, where we train our model using different proportions of the LSUN-B-ADM subset of DiGEN dataset and subsequently conduct cross-dataset evaluations. Remarkably, our method demonstrates robust performance and notable generalization capabilities, even when trained with only 25% of the training data. This resilience is largely attributed to our innovative multi-layer feature adaptation strategy. By optimizing the foundational model's (i.e., CLIP) generalization capabilities, our method effectively identifies universal Deepfake features serving as reliable cues to differentiate between real and fake. This is particularly useful under the conditions of reduced data availability, underscoring the efficiency of our approach in learning from limited data.

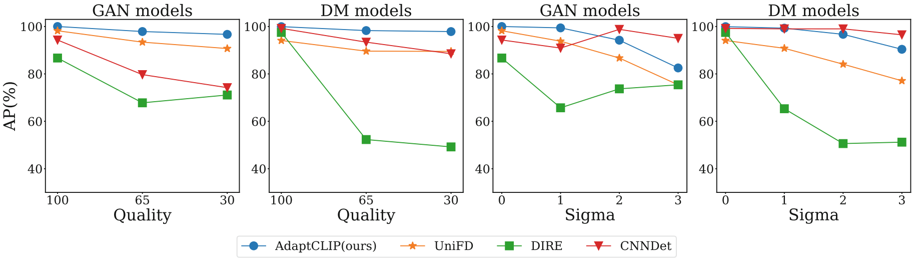


Fig. 5. Robustness against unseen perturbations. The left columns show robustness to JPEG compression, while the right columns display robustness to Gaussian blur. Performance is reported separately for GANs and DMs.

4.5 Robustness Against Unseen Perturbations.

In real-world applications, images are often subject to a variety of post-processing techniques, which underscore the importance of developing robust detectors against these unseen perturbations. To assess the model robustness, we examine the performance of detectors against two common types of image perturbations, JPEG compression ($quality = 100, 65, 30$) and Gaussian blur ($\sigma = 1, 2, 3$). We explore the robustness of our retrained baseline models UniFD [23], DIRE [36], CNNDet [35] and the proposed AdaptCLIP. The results are shown in Fig. 5. Our findings reveal that AdaptCLIP exhibits remarkable performance, maintaining its efficacy without significant degradation even when confronted with these real-world post-processing operations. This robustness can be attributed to the utilization of the multi-layer feature adaptation strategy that leverages the comprehensive feature set of the foundation model, effectively mitigating the impact of perturbations.

5 Conclusion

In this paper, we introduce an innovative adaptation mechanism tailored for foundation models (FM) (i.e., CLIP [26]), aiming at creating a generalized image-based Deepfake detector capable of identifying AI-generated images from both GANs and DMs. Additionally, we present the DiGEN dataset, a diverse and challenging dataset (i.e., it includes indoors, face, natural images from seven GANs, nine DMs) which can effectively complement the prior datasets. Our experimental evaluations reveal a critical insight: conventional detectors often struggle to accurately identify images generated by DMs. However, the extensive evaluation results demonstrate a remarkable ability of the proposed approach to generalize well to previously unseen Deepfakes. This capability also marks a substantial progression in the realm of generalized image-based Deepfake detection. By bridging advancements in generative AI with parallel developments in detection techniques, we envision a balanced progression in the field, ensuring both innovation and safety in generative AI.

Acknowledgements. This research is supported by National Science and Technology Council, Taiwan (R.O.C), under the grant number of NSTC-112-2634-F-002-006, NSTC-112-2222-E-001-001-MY2, NSTC-113-2634-F-001-002-MBK, NSTC-112-2218-E-011-012, NSTC-111-2221-E-011-128-MY3 and Academia Sinica under the grant number of AS-CDA-110-M09.

References

1. Bird, J.J., Lotfi, A.: Cifake: Image classification and explainable identification of ai-generated synthetic images. IEEE Access (2024)
2. Choi, J., Lee, J., Shin, C., Kim, S., Kim, H., Yoon, S.: Perception prioritized training of diffusion models. In: CVPR (2022)
3. Corvi, R., Cozzolino, D., Poggi, G., Nagano, K., Verdoliva, L.: Intriguing properties of synthetic images: from generative adversarial networks to diffusion models. In: CVPR (2023)
4. Corvi, R., Cozzolino, D., Zingarini, G., Poggi, G., Nagano, K., Verdoliva, L.: On the detection of synthetic images generated by diffusion models. In: ICASSP (2023)
5. Cozzolino, D., Poggi, G., Corvi, R., Nießner, M., Verdoliva, L.: Raising the bar of ai-generated image detection with clip. arXiv preprint arXiv:2312.00195 (2023)
6. Dhariwal, P., Nichol, A.: Diffusion models beat gans on image synthesis. NeurIPS (2021)
7. Epstein, D.C., Jain, I., Wang, O., Zhang, R.: Online detection of ai-generated images. In: ICCV (2023)
8. Gao, P., Geng, S., Zhang, R., Ma, T., Fang, R., Zhang, Y., Li, H., Qiao, Y.: Clip-adapter: Better vision-language models with feature adapters. IJCV (2024)

9. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. NeurIPS **27** (2014)
10. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. NeurIPS (2020)
11. Karras, T., Aila, T., Laine, S., Lehtinen, J.: Progressive growing of gans for improved quality, stability, and variation. arXiv preprint arXiv:1710.10196 (2017)
12. Karras, T., Aittala, M., Laine, S., Härkönen, E., Hellsten, J., Lehtinen, J., Aila, T.: Alias-free generative adversarial networks. NeurIPS (2021)
13. Karras, T., Laine, S., Aila, T.: A style-based generator architecture for generative adversarial networks. In: CVPR (2019)
14. Karras, T., Laine, S., Aittala, M., Hellsten: Analyzing and improving the image quality of stylegan. In: CVPR (2020)
15. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: ECCV (2014)
16. Lin, Z., Geng, S., Zhang, R., Gao, P., De Melo, G., Wang, X., Dai, J., Qiao, Y., Li, H.: Frozen clip models are efficient video learners. In: ECCV (2022)
17. Liu, B., Yang, F., Bi, X., Xiao, B., Li, W., Gao, X.: Detecting generated images by real images. In: ECCV (2022)
18. Liu, L., Ren, Y., Lin, Z., Zhao, Z.: Pseudo numerical methods for diffusion models on manifolds. arXiv preprint arXiv:2202.09778 (2022)
19. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101 (2017)
20. Luo, Y., Zhang, Y., Yan, J., Liu, W.: Generalizing face forgery detection with high-frequency features. In: CVPR (2021)
21. Mandelli, S., Bonettini, N., Bestagini, P., Tubaro, S.: Detecting gan-generated images by orthogonal training of multiple cnns. In: ICIP. IEEE (2022)
22. Nichol, A.Q., Dhariwal, P.: Improved denoising diffusion probabilistic models. In: ICML (2021)
23. Ojha, U., Li, Y., Lee, Y.J.: Towards universal fake image detectors that generalize across generative models. In: CVPR (2023)
24. Podell, D., English, Z., Lacey, K., Blattmann, A., Dockhorn, T., Müller, J., Penna, J., Rombach, R.: Sdxl: Improving latent diffusion models for high-resolution image synthesis. arXiv preprint arXiv:2307.01952 (2023)
25. Qian, Y., Yin, G., Sheng, L., Chen, Z., Shao, J.: Thinking in frequency: Face forgery detection by mining frequency-aware clues. In: ECCV (2020)
26. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: ICML (2021)
27. Ricker, J., Damm, S., Holz, T., Fischer, A.: Towards the detection of diffusion model deepfakes. arXiv preprint arXiv:2210.14571 (2022)
28. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: CVPR (2022)
29. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: CVPR (2022)
30. Sauer, A., Chitta, K., Müller, J., Geiger, A.: Projected gans converge faster. NeurIPS (2021)
31. Sha, Z., Li, Z., Yu, N., Zhang, Y.: De-fake: Detection and attribution of fake images generated by text-to-image generation models. In: CCS (2023)
32. Shiohara, K., Yamasaki, T.: Detecting deepfakes with self-blended images. In: CVPR (2022)

33. Tan, C., Zhao, Y., Wei, S., Gu, G., Liu, P., Wei, Y.: Frequency-aware deepfake detection: Improving generalizability through frequency space domain learning. In: AAAI (2024)
34. Tan, M., Le, Q.: Efficientnet: Rethinking model scaling for convolutional neural networks. In: ICML (2019)
35. Wang, S.Y., Wang, O., Zhang, R., Owens, A., Efros, A.A.: Cnn-generated images are surprisingly easy to spot... for now. In: CVPR (2020)
36. Wang, Z., Bao, J., Zhou, W., Wang, W., Hu, H., Chen, H., Li, H.: Dire for diffusion-generated image detection. In: ICCV (2023)
37. Wang, Z., Zheng, H., He, P., Chen, W., Zhou, M.: Diffusion-gan: Training gans with diffusion. arXiv preprint arXiv:2206.02262 (2022)
38. Yu, F., Seff, A., Zhang, Y., Song, S., Funkhouser, T., Xiao, J.: Lsun: Construction of a large-scale image dataset using deep learning with humans in the loop. arXiv preprint arXiv:1506.03365 (2015)
39. Zhang, L., Xu, Z., Barnes, C., Zhou, Y., Liu, Q., Zhang, H., Amirghodsi, S., Lin, Z., Shechtman, E., Shi, J.: Perceptual artifacts localization for image synthesis tasks. In: ICCV (2023)
40. Zhao, T., Xu, X., Xu, M., Ding, H., Xiong, Y., Xia, W.: Learning self-consistency for deepfake detection. In: ICCV (2021)

Audio Deepfake Detection: A Continual Approach with Feature Distillation and Dynamic Class Rebalancing

Taiba Majid Wani(✉) and Irene Amerini

Sapienza University of Rome, Rome, Italy
{majid,amerini}@diag.uniroma1.it
https://www.uniroma1.it/

Abstract. In an era where digital authenticity is frequently compromised by sophisticated synthetic audio technologies, ensuring the integrity of digital media is crucial. This paper addresses the critical challenges of catastrophic forgetting and incremental learning within the domain of audio deepfake detection. We introduce a novel methodology that synergistically combines the discriminative feature extraction capabilities of SincNet with the computational efficiency of LightCNN. Our approach is further augmented by integrating Feature Distillation and Dynamic Class Rebalancing, enhancing the model's adaptability across evolving deepfake threats while maintaining high accuracy on previously encountered data. The models were tested using the ASVspoof 2015, ASVspoof 2019, and FoR datasets, demonstrating significant improvements in detecting audio deepfakes with reduced computational overhead. Our results illustrate that the proposed model not only effectively counters the issue of catastrophic forgetting but also exhibits superior adaptability through dynamic class rebalancing and feature distillation techniques.

Keywords: Audio deepfakes · Continual learning · Catastrophic forgetting · Feature Distillation · Dynamic Class Rebalancing

1 Introduction

The rapidly evolving landscape of audio deepfake detection has become a critical area of research, with the advent of sophisticated speech synthesis and voice conversion technologies. These advancements have led to the generation of highly realistic audio deepfakes, posing significant threats to individual privacy, security, and the integrity of information. Consequently, the imperative for the establishment of robust and adaptable detection mechanisms has been accentuated. The inception of challenges such as the ASVspoof [1] and the Audio Deep Synthesis Detection (ADD) [2] challenge has propelled the field forward, showcasing the efficacy of deep neural networks in identifying audio deepfakes. These competitions have highlighted the crucial role of innovative architectures and feature

A. Antonacopoulos et al. (Eds.): ICPR 2024, LNCS 15321, pp. 211–227, 2025.
https://doi.org/10.1007/978-3-031-78305-0_14

extraction techniques in enhancing detection performance. However, the continual emergence of new deepfake variants presents a formidable challenge, as existing models often struggle to maintain their effectiveness against previously unseen attacks [3].

Convolutional Neural Networks (CNNs) have been a cornerstone in the area of audio deepfake detection systems, owing to their ability to extract hierarchical features from spectrograms and raw audio signals [4]. The adaptability and depth of CNNs make them suitable for capturing the nuances of genuine and fake audio, thus providing a solid foundation for classification. Despite their strengths, CNNs face limitations such as vulnerability to adversarial attacks, difficulty in handling audio's temporal dynamics, and catastrophic forgetting, where they lose the ability to detect older threats as they adapt to new ones [5].

Continual learning emerges as a promising solution to the challenges of audio deepfake detection, enabling models to incrementally learn from new data while retaining knowledge of past learnings [16]. This approach is crucial in addressing the dynamic nature of audio deepfake threats, where the constant evolution of attack methods necessitates adaptable models that can evolve over time without succumbing to catastrophic forgetting. By employing continual learning strategies, audio deepfake detection systems can maintain their effectiveness against new and evolving threats, ensuring a higher level of security and integrity in digital scenarios. Regularization-based continual learning methods offer an alternative approach that eliminates the need for retaining previous data [10]. These methods impose constraints on the model's parameters to prevent significant deviations from previously learned weights, thereby preserving the knowledge acquired from older tasks. Among the regularization-based techniques, the Detecting Fake Without Forgetting (DFWF) framework has been specifically designed for audio deepfake detection [10]. DFWF employs a regularization term that penalizes changes to crucial parameters responsible for detecting previously encountered deepfakes, effectively reducing the forgetting of past knowledge. While DFWF helps mitigate forgetting, it can struggle with new attack types due to the constraints it imposes on the model, limiting adaptability. In contrast, fine-tuning allows for better performance on new tasks by retraining the model without constraints, but this approach risks increased forgetting of previous tasks.

To address the challenges of catastrophic forgetting and incremental learning in audio deepfake detection, we propose a novel methodology that combines the strengths of SincNet [11] and LightCNN [12] with Feature Distillation (FD) and Dynamic Class Rebalancing (DCR), utilizing the ASVspoof 2015 [13], ASVspoof 2019 [14], and FoR [15] datasets for evaluation. In our framework, SincNet, renowned for its rich and discriminative feature representations through parameterized sinc functions that capture essential speech characteristics, serves as the teacher model. LightCNN, recognized for its lightweight and efficient architecture, acts as the student model, inheriting distilled knowledge from SincNet to learn crucial features necessary for accurate audio deepfake detection without the computational overhead. DCR further enhances our approach by adapting the learning strategy to address imbalanced and evolving data distributions, ensuring a balanced focus between preserving knowledge of previously encountered deepfakes and adapting to new ones. This integrated methodology effectively addresses catastrophic forgetting and ensures continuous learning in the face of evolving audio deepfake threats, as demonstrated by our experiments on the ASVspoof 2015, ASVspoof 2019, and FoR datasets.

2 Related Works

Recent advancements in continual learning approaches have significantly contributed to the field of audio deepfake detection, particularly in the context of the ASVspoof dataset. Continual learning aims to address the challenge of learning new tasks without forgetting previously acquired knowledge, which is crucial for adapting to evolving audio deepfake techniques. In this section, we review recent works that have explored various continual learning strategies and other deep learning architectures for audio deepfake detection using the ASVspoof datasets and FoR dataset.

H. Ma et. al., [16], introduced Detecting Fake Without Forgetting (DFWF), a continual learning approach for fake audio detection using the ASVspoof 2019 dataset. DFWF combined Learning Without Forgetting (LwF) with a Positive Sample Alignment (PSA) constraint and used LFCC features and a Light Convolutional Neural Network (LCNN) for classification. Evaluated with Equal Error Rate (EER) metrics, DFWF showed significant improvements over fine-tuning in sequential training tasks and offered a faster alternative to multi-condition training without needing previous data access.

X. Zhang et. al., [17] presented Radian Weight Modification (RWM), a continual learning approach that categorized classes based on feature distribution similarities and optimized gradient modification directions using a self-attention mechanism. The Wav2vec 2.0 model and a self-attention convolutional neural network (S-CNN) were employed for feature extraction and classification, respectively. Performance was evaluated on ASVspoof 2015, ASVspoof 2019 and In-the-Wild dataset [18] using the Equal Error Rate (EER), with RWM demonstrating superior ability in mitigating forgetting and acquiring new knowledge compared to other continual learning methods. In another work, X. Zhang et. al., [19], introduced Regularized Adaptive Weight Modification (RAWM), that adaptively modified the weight direction based on the ratio of genuine to fake utterances and incorporated a regularization constraint to preserve the old feature distribution of genuine audio. The approach effectively learned new spoofing attacks incrementally and mitigated catastrophic forgetting.

P. Kawa et. al., [20] investigated the robustness of audio DeepFake detection models against adversarial attacks and introduced an adaptive adversarial training method to enhance their robustness. Three models (LCNN, RawNet3, and SpecRNet) were evaluated on a combined dataset from ASVspoof2021 (DF subset), FakeAVCeleb, and WaveFake. The adaptive training method improved the robustness of the LCNN model, reducing the Equal Error Rate (EER) from 0.7870 to 0.1247 and showing improved performance in the transferability benchmark. This approach demonstrated the effectiveness of adaptive training in defending against adversarial attacks in deepfake detection.

J. Khochare et. al., [21] presented a deep learning framework for audio deepfake detection using the Fake or Real (FoR) dataset, which contained data generated by the latest text-to-speech models. Two approaches were adopted: a feature-based approach using machine learning algorithms (SVM, LGBM, XGBoost, KNN, and RF) with various spectral features (Mean Square Energy, Chroma Features, Spectral Centroid, Spectral Bandwidth, Spectral Rolloff, Zero Crossing Rate, and MFCCs) and an image-based approach using deep learning algorithms (Temporal Convolutional Network (TCN) and Spatial Transformer Network (STN)) with mel-spectrograms as input. TCN model outperformed machine learning algorithms and STN with a test accuracy of 92%, demonstrating the effectiveness of deep learning algorithms, particularly TCN, in audio deepfake detection.

Building on the state-of-the-art works in continual learning for audio deepfake detection, we propose a novel methodology that leverages the strengths of SincNet and LightCNN, complemented by Feature Distillation and Dynamic Class Rebalancing, to address the challenges of catastrophic forgetting and incremental learning. In our framework, SincNet serves as the teacher model, providing rich feature representations for audio data, while LightCNN acts as the student model, learning efficiently from the distilled knowledge of SincNet. Feature Distillation ensures that LightCNN inherits the comprehensive knowledge captured by SincNet, enabling it to learn essential features for accurate audio deepfake detection. Dynamic Class Rebalancing adapts the learning strategy based on the similarity of class features across tasks, maintaining a balance between preserving old knowledge and acquiring new information. This fine-tuned learning process prevents catastrophic forgetting, allowing the model to continuously learn from new data without losing previously acquired knowledge, thus offering a robust and adaptable solution for audio deepfake detection in a continual learning approach.

3 Proposed Methodology

In this study, we present a novel approach for audio deepfake detection that combines SincNet as a teacher model with LightCNN as a student model within a continual learning framework. The process begins with data collection from three datasets: ASVspoof 2015 (A), ASVspoof 2019 (B1), and FoR (for-original (B2), for-normalization (B3), for-2second (B4), and for-rerecording (B5). This is followed by preprocessing steps, including noise reduction, normalization, and silence removal, to prepare the data. SincNet, the teacher model, is trained on this preprocessed data to extract rich feature representations. These representations are then distilled into LightCNN through Feature Distillation (FD), enabling LightCNN to leverage these detailed features along with the MFCC and

LFCC features extracted from the preprocessed data while maintaining computational efficiency. LightCNN is incrementally trained with the distilled knowledge, and Dynamic Class Rebalancing (DCR) is employed to adjust the learning process based on class similarity, preventing catastrophic forgetting while learning new data. The model's performance is continually evaluated, and necessary adaptations are made to ensure optimal performance. This cyclical process, with continual updates as new datasets are introduced, ensures the model remains robust and accurate over time. The proposed methodology for audio deepfake detection is illustrated in Figure. 1.

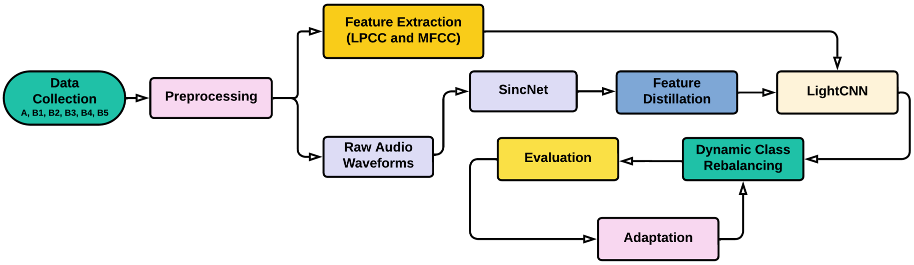


Fig. 1. Proposed pipeline for the Audio Deepfake Detection: including steps for data collection from various datasets, preprocessing, feature extraction using LPCC and MFCC, initial training of LightCNN, processing of raw audio waveforms through SincNet, Feature Distillation, LightCNN refinement, dynamic class rebalancing, evaluation, and adaptation for improved performance.

3.1 Data Preparation and Feature Extraction

The first step involves the collection of a comprehensive dataset comprising both real and fake audio samples. To ensure the robustness of our models, we include a diverse range of deepfake datasets, encompassing various techniques used for audio manipulation. The preprocessing phase includes several key steps, such as noise reduction to eliminate background noise that could obscure important audio features, normalization to ensure uniform amplitude levels across all samples, and silence removal to exclude non-informative segments of the audio. These preprocessing steps are essential for reducing variability in the data and improving the model's ability to focus on relevant audio characteristics.

Feature extraction is a pivotal component of the methodology, as it transforms raw audio data into a format that is more suitable for analysis by our deepfake detection models. For the SincNet model, we utilize raw audio waveforms directly, leveraging the model's capability to learn effective filterbanks from the data. For LightCNN, we extract two types features, Mel-Frequency Cepstral Coefficients (MFCCs) and Linear Frequency Cepstral Coefficients (LFCCs), which provide a compact representation of the audio's spectral properties.

Finally, we split the dataset into training, development, and evaluation, ensuring that each set contains a stratified distribution of classes. This division is crucial for maintaining balance between real and fake samples in all subsets of the data, which helps prevent bias in the model training and evaluation processes.

3.2 Model Setup

In the second step of the methodology, we establish the foundational models for our audio deepfake detection system. This involves setting up both the teacher model, which provides a source of rich audio feature representations, and the student model, which learns from the teacher while adapting to new data in a continual learning framework.

Teacher Model (SincNet). We selected SincNet as the teacher model due to its specialized design for processing raw audio waveforms. SincNet employs parameterized sinc functions to create learnable band-pass filters, which efficiently extract fine-grained speech features directly from the audio input [11]. This capability makes SincNet particularly well-suited for tasks involving speech and audio analysis, such as deepfake detection.To train the SincNet model, we use the training set prepared in the step 1. The objective of the training is to minimize a loss function, typically the cross-entropy loss, which is defined as:

$$L_{\text{SincNet}} = -\sum_{i=1}^{N}\sum_{c=1}^{C} y_{i,c} \log(\hat{y}_{i,c}^{\text{SincNet}})$$

where N is the number of samples, C is the number of classes, $y_{i,c}$ is the true label, and $(\hat{y}_{i,c}^{\text{SincNet}})$ is the predicted probability by the SincNet model for class c of the $i - th$ sample.

Student Model (LightCNN). For the student model, we select LightCNN due to its lightweight architecture and efficiency in learning. It employs Max-Feature-Map (MFM) activations instead of traditional ReLU activations, which helps in reducing the model size while maintaining performance. Compared to SincNet, LightCNN offers efficient runtime and memory savings. Due to its simpler architecture, LightCNN reduces memory usage by approximately 30-50% and improve inference speed by 20-40%. These efficiencies make it particularly suitable for the applications where computational resources are limited. The LightCNN model undergoes incremental training on new tasks, with the initial training phase involving knowledge distillation from the SincNet teacher model. This process allows LightCNN to inherit the rich audio feature representations learned by SincNet, providing a strong foundation for its subsequent adaptation to new deepfake types.

3.3 Initial Training on Base Task

In Step 3 , we focus on the initial training of the LightCNN student model and the transfer of knowledge from the SincNet teacher model through feature distillation. ***Training LightCNN.***The LightCNN model is first trained on an initial dataset or task, which serves as the base task for the subsequent incremental learning process. This initial training phase is essential for establishing baseline performance and preparing the model for future adaptations. The training objective for LightCNN is typically to minimize a task-specific loss function, such as cross-entropy for classification tasks. The cross-entropy loss is defined as:

$$L_{\text{LightCNN}} = -\sum_{i=1}^{N}\sum_{c=1}^{C} y_{i,c}\log(\hat{y}_{i,c}^{\text{LightCNN}})$$

where N is the number of samples, C is the number of classes, $y_{i,c}$ is the true label, and $\hat{y}_{i,c}^{\text{LightCNN}}$ is the predicted probability by the LightCNN model for class c of the i-th sample. This loss function measures the discrepancy between the predicted class probabilities and the true labels, guiding the model to learn accurate representations of the audio data.

Feature Distillation. After the initial training of LightCNN, we employ feature distillation to transfer knowledge from the SincNet teacher model to the LightCNN student model. This process involves minimizing a distillation loss that encourages the student model to mimic the feature representations of the teacher model. The distillation loss can be defined as the mean squared error (MSE) between the features extracted by the two models:

$$L_{\text{distill}} = \frac{1}{N}\sum_{i=1}^{N} \|\phi(\mathbf{x}_i^{\text{LightCNN}}) - \phi(\mathbf{x}_i^{\text{SincNet}})\|^2$$

where $\phi(\cdot)$ represents the feature extractor function of the model, and $\mathbf{x}_i^{\text{LightCNN}}$ and $\mathbf{x}_i^{\text{SincNet}}$ are the features extracted by the LightCNN and SincNet models, respectively, for the i-th sample. By minimizing this loss, the LightCNN model learns to produce feature representations that closely resemble those of the SincNet model, effectively inheriting the teacher model's knowledge about the audio data.

3.4 Incremental Learning for New Tasks

In Step 4, we address the challenge of incremental learning for new tasks, enabling the LightCNN student model to adapt to emerging deepfake types while retaining knowledge from previous tasks. This step involves a combination of Dynamic Class Rebalancing (DCR) mechanism and Feature Distillation.

DCR Mechanism (Class Categorization). DCR is a mechanism that adjusts the training process based on the similarity of class features across different tasks. Classes with similar features across tasks are grouped together, while those with distinct features are placed in separate groups. Clustering algorithms or similarity metrics, such as cosine distance, is used for this purpose:

$$\text{cosine distance}(x, y) = 1 - \frac{x \cdot y}{\|x\|\|y\|}$$

By identifying classes with similar feature distributions, the model focuses on preserving knowledge for these classes while adapting more freely to classes with new or distinct features.

DCR Mechanism (Gradient Modification). Once classes are categorized, the DCR mechanism adjusts the gradient update strategy during training. For classes with similar features across tasks, the model minimizes gradient modification to prevent the loss of previously learned knowledge. This might include adopting a lower learning rate or implementing regularization methods for these classes. Conversely, for classes with dissimilar features, the model updates the weights by modifying the gradient direction to be orthogonal to the previous data plane. This orthogonal modification involves projecting the gradient onto the subspace that is orthogonal to the gradient direction associated with previously learned tasks. This is mathematically expressed as:

$$P_{\perp}(\nabla L) = \nabla L - \frac{\nabla L \cdot \nabla L_{\text{prev}}}{\|\nabla L_{\text{prev}}\|^2} \nabla L_{\text{prev}}$$

∇L is the current gradient and ∇L_{prev} is the gradient from previous tasks. By ensuring the new gradient is orthogonal, the model can learn new information without interfering with the knowledge acquired from earlier tasks. The detailed algorithm for DCR is outlined in Algorithm 1.

The DCR mechanism employs a directional adjustment of gradient vectors, which is particularly effective in rebalancing the learning process for classes with imbalanced feature distributions. By orienting gradient vectors towards the decision boundary that separates similar from dissimilar classes, DCR dynamically corrects the learning trajectory to prioritize underrepresented classes. This approach enhances the model's ability to adapt to new data while maintaining performance on previously learned tasks. This method contrasts with the Radian Weight Modification (RWM) approach [17], which adjusts gradients based on a fixed radial angle. The RWM method, while effective, is less responsive to evolving data topology compared to DCR, potentially limiting its ability to enhance class discriminability in the learned feature space. Fig. 2 illustrates these conceptual differences in continual learning strategies. Fig 2(a) demonstrates the Dynamic Class Rebalancing (DCR) method, with blue and red vectors indicating the gradients for classes with similar and dissimilar features, respectively. The minimal adjustment of similar class gradients and the orthogonal modification of dissimilar class gradients help maintain a balanced focus between old and new information. Fig 2(b) shows the Radian Weight Modification (RWM) method, where the green vector represents the original gradient and the purple vector represents its RWM-adjusted direction. The fixed radial angle adjustment in RWM is less adaptive compared to the dynamic nature of DCR.

Feature Distillation (FD Technique). Throughout the incremental learning process, the FD technique continues to transfer knowledge from the SincNet teacher model to the LightCNN student model. This is achieved by minimizing the distillation loss, which encourages the student model to align its feature representations with those of the teacher model. Feature distillation ensures

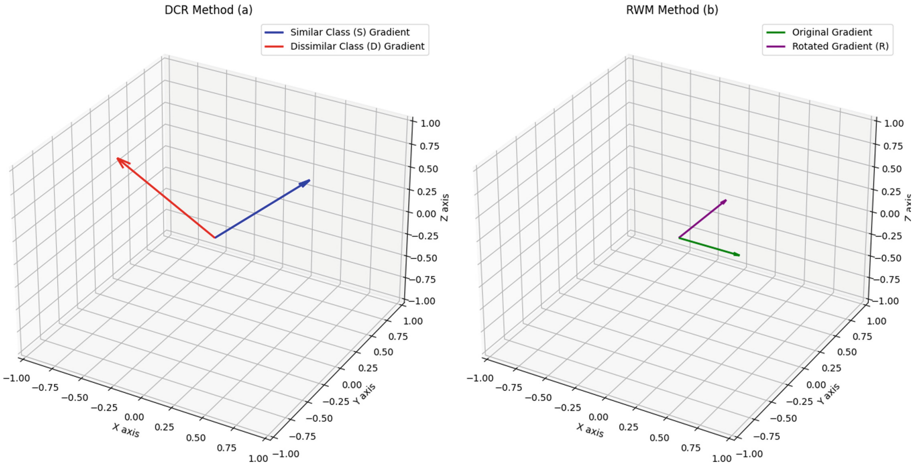


Fig. 2. The conceptual representations of Continual Learning strategies. (a) The DCR method shows how gradients are adjusted based on class feature similarities across tasks. The blue vector represents the Similar Class (S) Gradient and the red vector represents the Dissimilar Class (D) Gradient. (b) The RWM method illustrates the original gradient (green vector) and its rotated counterpart (purple vector) after applying a self-adaptive rotation. (Color figure online)

that the student model retains the rich and nuanced understanding of audio data provided by the teacher model, even as it learns from new tasks.

3.5 Evaluation and Adaptation

After training the LightCNN model on each new task, we perform a comprehensive evaluation to assess its performance. This evaluation includes measuring detection accuracy on the test sets of all seen tasks and EER%, which provides insight into the model's ability to correctly identify real and fake audio samples. Additionally, we evaluate the extent of catastrophic forgetting by comparing the model's performance on previous tasks before and after training on the new task. This helps us understand how well the model retains knowledge from earlier tasks while learning new information. Based on the evaluation results, we adapt the LightCNN model to ensure optimal performance. Model adaptation involves fine-tuning the model parameters, adjusting the learning rate, or revisiting the feature distillation process to reinforce the knowledge transfer from the SincNet teacher model. The goal of this adaptation is to enhance the model's performance on new tasks while preserving its accuracy on previously learned tasks. ***Continual Learning Cycle.*** The process of incremental learning, evaluation, and adaptation forms a continual learning cycle that is repeated for each new task. By continually updating the LightCNN student model, we enable it to learn from new data without forgetting the knowledge acquired from previous tasks. This cycle is essential for keeping the model up-to-date and effective in

Algorithm 1 Dynamic Class Rebalancing (DCR) for Continual Learning

```
Require: Training data from multiple tasks, $\{D_t\}$; Learning rate, $\gamma$; Regularization
    constant, $\lambda$; Task count, $t$; Feature extractor function of the teacher model, $\Phi_{\text{teach}}$
Initialize LightCNN model weights, $W_0$
for each task $t$ in $\{D_t\}$ do
    for each class $c$ in task $t$ do
        Compute class mean feature vector using the teacher model, $\mu_c = \Phi_{\text{teach}}(\text{mean}(D_t[c]))$
    end for
    Compute pairwise cosine similarities for all classes, $S = \{s(c_i, c_j)\}$
    Categorize classes into groups based on $S$:
    $G_{\text{sim}} = \{c \mid \forall c_i \text{ in task } t, \forall c_j \text{ in task } t, s(c_i, c_j) > \text{similarity\_threshold}\}$
    $G_{\text{diff}} = \{c \mid c \notin G_{\text{sim}}\}$
    for each training iteration $i$ do
        Compute the batch loss, $L$, and its gradient, $\nabla L$
        if class of batch in $G_{\text{sim}}$ then
            Perform regularized update on $W$: $W_{i+1} = W_i - \gamma \nabla L + \lambda R(W_i)$
        else
            Compute orthogonal projection of $\nabla L$ onto subspace orthogonal to $G_{\text{sim}}$, $P_{\perp}(\nabla L)$
            Perform update using orthogonal gradient: $W_{i+1} = W_i - \gamma P_{\perp}(\nabla L)$
        end if
    end for
    Evaluate and adapt model based on the validation set for task $t$
end for
return adapted LightCNN model weights, $W_t$
```

detecting evolving deepfake techniques. At the end of the learning process, we conduct a final comprehensive evaluation of the LightCNN student model. This evaluation assesses the model's overall performance across all tasks, providing a comprehensive view of its capabilities in audio deepfake detection.

4 Experimental Setup and Results

4.1 Datasets

We have employed a structured approach to training, evaluation, and development using the ASVspoof 2015, ASVspoof 2019, and FoR datasets for audio deepfake detection in a continual learning framework. We begin with the ASVspoof 2015 dataset, which serves as the initial training set (A). This dataset lays the foundation for our model, enabling it to learn the essential features of audio deepfakes. We divide this dataset into training, development, and evaluation subsets to effectively train and validate the model's performance. Once the model is trained on the ASVspoof 2015 dataset, we incrementally introduce the ASVspoof 2019 dataset (LA senario) (B1) to the training process. This step is crucial for continual learning, as it allows the model to update its knowledge with new data while retaining the information learned from the ASVspoof 2015 dataset. We use a subset of the ASVspoof 2019 dataset for training and another subset for evaluation, ensuring a balanced approach to learning and validation.

Following the incorporation of the ASVspoof 2019 dataset, we further extend the model's learning with the FoR dataset. FoR dataset has four different versions, for-original (B2), for-normalization (B3), for-2seconds (B4) and for-rerecording (B5). This dataset adds diversity and complexity to the training process, challenging the model's ability to adapt and learn incrementally. Similar to the previous datasets, we divide B2, B3, B4 and B5 datasets into training, development and evaluation subsets, allowing for continuous assessment of the model's performance.

Throughout the training and incremental learning process, we consistently evaluate the model on separate evaluation sets that are not used during training. This continuous evaluation is critical for monitoring the model's ability to generalize to unseen data. Additionally, we utilize development sets from each dataset to fine-tune hyperparameters and make informed decisions regarding the model's architecture. This structured approach ensures that our model undergoes a comprehensive and effective continual learning process, addressing the challenges of catastrophic forgetting and incremental learning in audio deepfake detection. Table 1 showcases the comprehensive statistics of the datasets used.

Table 1. Comprehensive statistics of datasets with real and fake utterances.

Datasets	A		B1		B2		B3		B4		B5	
Phases	Real	Fake	Real	Fake	Real	Fake	Real	Fake	Real	Fake	Real	Fake
Train	3,750	12,625	2,580	22,800	39,254	57,060	26,927	26,941	5,104	5,104	6,978	6,978
Dev	3,497	49,875	2,548	22,296	14,265	24,262	5,398	5,400	1,143	1,101	1,413	1,413
Eval	9,404	1,84,000	7,355	63,882	24,215	36,485	2,370	2,264	408	408	544	544

4.2 Experimental Setup

We deploy a two-tier architecture integrating SincNet as the teacher model for its feature extraction capabilities from raw audio data and LightCNN as the student model, optimized for efficiency and performance. We utilize the Adam optimizer for its adaptive learning rate capabilities, with a batch size of 4 and an initial learning rate of 0.0001, which is dynamically adjusted based on validation performance. To enhance the precision and robustness of our model, we have implemented a thorough training strategy, extending over 100 epochs. For Feature Distillation, weight parameters are tuned to optimize knowledge transfer from SincNet to LightCNN and DCR parameters are adaptively set based on evolving data distributions and class similarity metrics, ensuring effective learning from both balanced and imbalanced data scenarios.

4.3 Results and Analysis

The initial training phase of our model utilized the ASVspoof 2015 dataset (A). Table 2, depicts the EER% for the baseline model across various evaluation sets. It is important to note that these results represent the performance of the base model trained only on the initial dataset without incremental learning. As shown in Table 2, while the model demonstrates good performance on the A, a decline in detection accuracy is observed when applied to other datasets ($B1 \rightarrow B5$). This clearly illustrates the critical challenge of generalization and the need for incremental learning techniques to maintain performance across diverse audio datasets.

Table 2. Baseline Equal Error Rate (EER)% across various evaluation sets.

Model	A	B1	B2	B3	B4	B5
Baseline	1.45	10.56	45.98	35.16	64.78	75.89

Performance with SincNet directly We conducted experiments involving only SincNet with DCR and FD techniques. Table 3 shows the EER% for using SincNet directly with different η values across various datasets (A, B1, B2, B3, B4, B5). The parameter η plays a pivotal role in mediating the balance between FD and DCR. Specifically, η regulates the extent to which the model prioritizes the retention of previously distilled features over adapting to novel class distributions encountered in new datasets.

Table 3. EER% for using SincNet directly with different η values across all datasets.

η	A	B1	B2	B3	B4	B5
0	3.13	4.56	5.50	7.04	7.89	9.06
0.5	2.35	4.45	3.98	5.67	5.21	7.57
1	3.51	4.31	5.11	4.80	6.72	8.61

Table 3 indicates that using SincNet directly in the continual learning process with DCR and FD techniques leads to performance degradation from dataset $A \rightarrow B5$, regardless of the η value, as the model encounters more complex and diverse datasets.

Performance Using the Proposed Method Table 4 presents the EER% across evaluation sets for the proposed method under varying η values. An elevated value of η emphasizes the conservation of distinct, discriminative feature representations extracted by SincNet and subsequently distilled into LightCNN.

The varied η values reveal insights into the model's learning and forgetting behaviors, as shown in Table 4. A balanced η value (0.50) enhances the model's performance on both familiar and novel datasets. The experimental results, particularly in the sequential training from ($A \rightarrow B5$), underscore our method's capability to minimize performance degradation across datasets, with $\eta = 0.50$ leading to the lowest EERs.

Table 4. EER% under different η values. (a), (b), (c), (d) and (e) detail the training from dataset $A \rightarrow B_n$ dataset, with evaluations conducted on both A and the respective B_n dataset. (f) describes the training sequence from $A \rightarrow B1 \rightarrow B2 \rightarrow B3 \rightarrow B4 \rightarrow B5$, with evaluations performed across all datasets.

(a)

η	A	B1
0	1.41	0.63
0.5	**0.54**	**0.42**
1	1.32	0.54

(b)

η	A	B2
0	1.63	1.23
0.5	**0.94**	**1.02**
1	2.63	1.98

(c)

η	A	B3
0	2.18	1.84
0.5	**1.63**	**1.21**
1	2.25	2.41

(d)

η	A	B4
0	2.61	3.4
0.5	**2.23**	**2.19**
1	3.65	2.32

(e)

η	A	B5
0	3.46	2.43
0.5	**1.90**	**1.46**
1	2.83	3.21

(f)

η	A	B1	B2	B3	B4	B5
0	1.94	2.41	2.01	3.65	2.23	3.12
0.5	**1.21**	**0.98**	**1.46**	**1.23**	**1.84**	**2.22**
1	2.67	1.86	2.25	3.77	3.18	3.35

In Table 4(f), EER% represents the performance of the model after incremental learning. This includes evaluations on the new dataset as well as all previously seen datasets, ensuring that the model continues to retain knowledge from initial learning while adapting to new information. The results demonstrate the effectiveness of our continual learning approach, which mitigates catastrophic forgetting and maintains high detection accuracy across all tasks.

Comparing Table 4(f) and Table 3 shows that the proposed method consistently outperforms using SincNet directly across all datasets and η values. The proposed method, especially with $\eta = 0.5$, demonstrates better adaptability and robustness in audio deepfake detection, while using SincNet directly results in higher EERs and less effective handling of evolving threats. This highlights the effectiveness of integrating SincNet with LightCNN, FD, and DCR in our approach.

4.4 Evaluation on ASVspoof 2021 dataset

We also conducted experiments where the model was trained using an incremental learning approach. The training sequence followed the order of $A \rightarrow B1 \rightarrow B2 \rightarrow B3 \rightarrow B4 \rightarrow B5$ and was subsequently tested on the ASVspoof 2021 dataset [22]. This approach ensures that the model's performance, including its generalizability, is assessed against the latest benchmark in audio deepfake detection, which presents new challenges and variations in deepfake audio.

Table 5. EER% for incremental learning performance tested on ASVspoof 2021.

η	Training Datasets	Test Dataset	EER (%)
0	$A \rightarrow B1 \rightarrow B2 \rightarrow B3 \rightarrow B4 \rightarrow B5$	ASVspoof 2021	3.04
0.5	$A \rightarrow B1 \rightarrow B2 \rightarrow B3 \rightarrow B4 \rightarrow B5$	ASVspoof 2021	2.67
1	$A \rightarrow B1 \rightarrow B2 \rightarrow B3 \rightarrow B4 \rightarrow B5$	ASVspoof 2021	3.16

The results in Table 5 indicate that our proposed methodology effectively handles new variations and challenges presented by the latest deepfake audio samples. The model achieves a competitive EER on the ASVspoof 2021 dataset, with the lowest EER observed at $\eta = 0.5$. This demonstrates the model's ability to generalize to unseen data while maintaining robust performance.

4.5 Ablation Study

We conducted the ablation study to evaluate the independent contributions of the FD and DCR components in our proposed methodology as shown in Table 6. The study revealed that the full incorporation of both FD and DCR consistently yielded the lowest EER, signifying their critical roles in enhancing detection accuracy. Specifically, removal of FD generally led to moderate increase in EER, while exclusion of DCR resulted in more substantial deteriorations, indicating DCR's significant impact on maintaining model robustness across varying deepfake scenarios. These results underscore the effectiveness of our integrated approach in reducing error rates and highlight the importance of each component in fortifying the model against evolving audio deepfake challenges.

4.6 Comparison with state-of-art methods

We have compared our Dynamic Class Rebalancing (DCR) approach against established state-of-the-art methods in the area of continual learning and audio deepfake detection, as illustrated in Table 7. The benchmarked methods include Elastic Weight Consolidation (EWC), Learning without Forgetting (LwF), DeepFake WaveForm (DFWF), Orthogonal Weight Modification (OWM), Copy Weight Regularization (CWR), and Gradient Descent Feature (GDF).

Table 6. EER % determined for the evaluation datasets of the ablation studies, (a), (b), (c), and (d) were trained on the training set according to the sequence $A \rightarrow B_n$ and evaluated using the evaluation sets for A and B_n. (e) was trained using a sequential training set structured as $A \rightarrow B_1 \rightarrow B_2 \rightarrow B_3 \rightarrow B_4 \rightarrow B_5$ and was assessed using the corresponding evaluation sets.

(a)

Method	A	B1
Proposed	**0.08**	**1.69**
-FD	2.21	1.93
-DCR	3.68	4.41

(b)

Method	A	B2
Proposed	**2.93**	**1.74**
-FD	4.31	3.29
-DCR	3.68	4.38

(c)

Method	A	B3
Proposed	**2.18**	**1.84**
-FD	7.81	6.32
-DCR	10.42	8.17

(d)

Method	A	B4
Proposed	**3.65**	**2.11**
-FD	5.12	3.87
-DCR	6.43	4.18

(e)

Method	A	B5
Proposed	**3.41**	**1.98**
-FD	7.13	6.43
-DCR	9.26	5.67

(f)

Method	A	B1	B2	B3	B4	B5
Proposed	**1.18**	**4.64**	**2.86**	**3.19**	**3.62**	**5.19**
-FD	3.47	6.82	7.54	5.32	8.55	6.72
-DCR	6.19	7.48	9.56	4.09	7.72	8.3

EWC and LwF have shown efficient adaptability, with EERs of 5.97% and 3.14% on dataset A, respectively. DFWF, while superior in mitigating forgetting, reveals limitations to new dataset conditions, demonstrated by a 2.21% EER on dataset A which then escalates in dataset B5. OWM, CWR, and GDF also perform well but show higher EERs compared to our proposed method in several datasets.

In contrast, our DCR methodology showcases better performance, significantly lowering the EER across all datasets and demonstrating remarkable adaptability and memory retention. Specifically, DCR records an EER of 1.21% on dataset A, outperforming EWC, LwF, DFWF, OWM, CWR, and GDF, and maintains the lowest EER across datasets B1 through B5. This exemplifies our method's efficiency in learning from new datasets while effectively preserving knowledge from previous datasets, with notable performance on dataset B1 with an EER of 0.98% and on dataset B5 with an EER of 2.22%.

5 Conclusion

In this paper, we presented a novel methodology combining SincNet's feature extraction with LightCNN's efficiency, enhanced by Feature Distillation and Dynamic Class Rebalancing (DCR), to address audio deepfake detection. Evaluated on challenging datasets like ASVspoof 2015, 2019, and FoR, our approach

Table 7. Comparative EER% between our proposed and existing methods ($\eta = 0.5$).

Methods	A	B1	B2	B3	B4	B5
EWC	5.97	4.23	3.18	3.45	4.72	3.01
LwF	3.14	2.46	4.19	4.81	3.10	3.46
DFWF	2.21	1.48	3.32	2.91	3.65	4.65
OWM	3.78	2.17	2.89	4.11	3.54	3.11
CWR	4.23	2.81	3.72	3.17	2.47	2.91
GDF	5.39	3.66	2.47	3.67	4.34	3.45
DCR(Ours)	**1.21**	**0.98**	**1.46**	**1.23**	**1.84**	**2.22**

significantly outperformed state-of-the-art methods by achieving lower Equal Error Rate (EER) percentages. The strategic use of η for balancing knowledge retention and adaptation to new data was key to our success, enabling our model to excel in detecting audio deepfakes while minimizing catastrophic forgetting. This advancement not only sets a new benchmark in the field but also underscores the potential of integrating advanced distillation and rebalancing techniques for enhanced detection accuracy in the ongoing battle against audio deepfakes. Future directions could explore the integration of generative adversarial networks (GANs) for generating more diverse training data.

Acknowledgements. This study has been partially supported by SERICS (PE00000014) under the MUR National Recovery and Resilience Plan funded by the European Union – NextGenerationEU and Sapienza University of Rome project 2022–2024 "EV2" (003 009 22).

References

1. Wu, Z., Yamagishi, J., Kinnunen, T., Hanilçi, C., Sahidullah, M., Sizov, A., Evans, N., Todisco, M., Delgado, H.: ASVspoof: the automatic speaker verification spoofing and countermeasures challenge. IEEE Journal of Selected Topics in Signal Processing **11**(4), 588–604 (2017)
2. J. Yi, R. Fu, J. Tao, S. Nie, H. Ma, C. Wang, T. Wang, Z. Tian, Y. Bai, C. Fan, et al., "Add 2022: the first audio deep synthesis detection challenge," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 9216–9220, 2022
3. Dixit, A., Kaur, N., Kingra, S.: Review of audio deepfake detection techniques: Issues and prospects. Expert. Syst. **40**(8), e13322 (2023)
4. T. M. Wani and I. Amerini, "Deepfakes audio detection leveraging audio spectrogram and convolutional neural networks," in *International Conference on Image Analysis and Processing*, pp. 156–167, 2023
5. Zhang, B., Tondi, B., Barni, M.: Adversarial examples for replay attacks against CNN-based face recognition with anti-spoofing capability. Comput. Vis. Image Underst. **197**, 102988 (2020)

6. H. Ma, J. Yi, J. Tao, Y. Bai, Z. Tian, and C. Wang, ”Continual learning for fake audio detection,” arXiv preprint arXiv:2104.07286, 2021
7. H. Shin, J. K. Lee, J. Kim, and J. Kim, ”Continual learning with deep generative replay,” *Advances in Neural Information Processing Systems*, vol. 30, 2017
8. Tadros, T., Krishnan, G.P., Ramyaa, R., Bazhenov, M.: Sleep-like unsupervised replay reduces catastrophic forgetting in artificial neural networks. Nat. Commun. **13**(1), 7742 (2022)
9. Y. Patel, S. Tanwar, R. Gupta, P. Bhattacharya, I. E. Davidson, R. Nyameko, S. Aluvala, and V. Vimal, ”Deepfake Generation and Detection: Case Study and Challenges,” *IEEE Access*, 2023
10. L. Wang, X. Zhang, H. Su, and J. Zhu, ”A comprehensive survey of continual learning: Theory, method and application,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024
11. M. Ravanelli and Y. Bengio, ”Speaker recognition from raw waveform with SincNet,” in *2018 IEEE Spoken Language Technology Workshop (SLT)*, pp. 1021–1028, 2018
12. C. Liu, J. Li, J. Duan, H. Shen, and H. Huang, ”LightCvT: Audio forgery detection via fusion of light CNN and transformer,” in *Proceedings of the 2021 10th International Conference on Computing and Pattern Recognition*, pp. 99–105, 2021
13. Z. Wu, J. Yamagishi, T. Kinnunen, C. Hanilçi, M. Sahidullah, A. Sizov, N. Evans, M. Todisco, and H. Delgado,”ASVspoof: the automatic speaker verification spoofing and countermeasures challenge,”*IEEE Journal of Selected Topics in Signal Processing*, vol. 11, no. 4, pp. 588–604, 2017.
14. M. Todisco, X. Wang, V. Vestman, Md. Sahidullah, H. Delgado, A. Nautsch, J. Yamagishi, N. Evans, T. Kinnunen, and K. A. Lee, ”ASVspoof 2019: Future horizons in spoofed and fake audio detection,” arXiv preprint arXiv:1904.05441, 2019
15. R. Reimao and V. Tzerpos, ”FOR: A dataset for synthetic speech detection,” in *2019 International Conference on Speech Technology and Human-Computer Dialogue (SpeD)*, pp. 1–10, 2019
16. H. Ma, J. Yi, J. Tao, Y. Bai, Z. Tian, and C. Wang, ”Continual learning for fake audio detection,” arXiv preprint arXiv:2104.07286, 2021
17. X. Zhang, J. Yi, C. Wang, C. Zhang, S. Zeng, and J. Tao, ”What to remember: Self-adaptive continual learning for audio deepfake detection,” arXiv preprint arXiv:2312.09651, 2023
18. N. M. Müller, P. Czempin, F. Dieckmann, A. Froghyar, and K. Böttinger, ”Does audio deepfake detection generalize?”, arXiv preprint arXiv:2203.16263, 2022
19. X. Zhang, J. Yi, J. Tao, C. Wang, and C. Yuan Zhang, ”Do you remember? Overcoming catastrophic forgetting for fake audio detection,” in *International Conference on Machine Learning*, pp. 41819–41831, 2023
20. P. Kawa, M. Plata, and P. Syga, ”Defense against adversarial attacks on audio deepfake detection,” arXiv preprint arXiv:2212.14597, 2022
21. J. Khochare, C. Joshi, B. Yenarkar, S. Suratkar, and F. Kazi, “A deep learning framework for audio deepfake detection,” *Arabian Journal for Science and Engineering*, pp. 1–12, 2021
22. Yamagishi, J., Wang, X., Todisco, M., Sahidullah, M., Patino, J., Nautsch, A., Liu, X., Lee, K. A., Kinnunen, T., Evans, N., et al. “ASVspoof 2021: accelerating progress in spoofed and deepfake speech detection,” arXiv preprint arXiv:2109.00537, 2021

Diffusion Models as a Representation Learner for Deepfake Image Detection

Rajjeshwar Ganguly(✉), Mamadou Dian Bah, and Mohamed Dahmane

Computer Research Institute of Montreal (CRIM), Montreal, Canada
rajjeshwar.ganguly@crim.ca,rajjeshwar.ganguly@umontreal.ca,
mamadou-dian.bah@crim.ca, mohamed.dahmane@crim.ca

Abstract. This paper explores leveraging representations of the data distribution learned by diffusion models to improve a downstream task of deepfake image detection. With the recent upsurge in the popularity of generative AI, it has become increasingly common to encounter disinformation in modalities such as language and text, images and speech. However, a significant portion of disinformative content can solely be attributed to deepfake images. Effective countermeasures in the past have relied upon classifying deepfake images based on spatial irregularities, inconsistencies in high frequency content and fingerprint matching with known residuals from popular deepfake generation models. However, as the technology behind deepfakes continues to advance, there is a growing need for robust detection methods and tools to ensure the integrity of visual information and mitigate the risks associated with the spread of misleading or malicious content. Thus, we investigate using diffusion-generated reconstructions and latent space inversion to enhance deepfake detection, adapting to the changing landscape of visual disinformation. We explore the feasibility of using diffusion generated reconstruction, diffusion generated latent space inversion and high frequency feature extraction for improving the performance of detecting deepfakes.

Keywords: Deepfake · Representation Learning · Diffusion Models

1 Introduction

Deepfake images are digitally manipulated media content created using advanced artificial intelligence techniques, particularly deep learning algorithms. These sophisticated tools enable the seamless insertion or substitution of faces and facial expressions in existing images or videos, making it increasingly challenging to discern between real and fabricated content. The necessity to identify deepfake images arises from the potential for misuse and deception, as these manipulations can be exploited to spread false information, damage reputations, or even influence public opinion. The creation of these deepfake images are usually facilitated by readily accessible off the shelf/opensource unconditional or text-to-image deep generative networks based on GANs [22] or Diffusion models [28]. Besides this, procedures such as image editing [3], inpainting [20] as

A. Antonacopoulos et al. (Eds.): ICPR 2024, LNCS 15321, pp. 228–241, 2025.
https://doi.org/10.1007/978-3-031-78305-0_15

well as image blending [15] are also some of the other techniques used to create counterfeit images by changing one or multiple aspects of some original image. Traditional approaches to identify such images typically comprise directly training discriminative deep learning architectures derived from CNNs [1]. Although robust for the vast majority of cases, this approach typically relies solely on identifying irregularities in the semantic representation of an image, such as edges, colors, textures etc. More recent, methods also invoke the use of augmentations in the training data to magnify and accentuate these irregularities [13], making this class of models more robust to slight changes in the patterns observed in a purely synthetic or modified image. Besides classifying the raw images directly through a classifier, there are other approaches that utilize Discrete Fourier Transforms (DFT) [25] or Discrete Cosine Transforms (DCT) [6] of the raw images as the input to make the model learn high frequency trends that might differ for synthetic images as opposed to real images. Another method for detecting deepfake generations involves the study of residual noise or 'fingerprints' from images and comparing them to a vector representation of similarly obtained fingerprints of popular synthetic image generators [35]. However, despite the progress made in this domain, synthetic images obtained from recent image generators incrementally have lower FID scores and higher likelihood [4] compared to their former counterparts. One intriguing avenue within this domain is the study of diffusion models [11], a class of architectures that uniquely captures the flow of information through neural networks. Unlike traditional feedforward architectures, diffusion models introduce a dynamic learning process. They do this by allowing information to iteratively diffuse through the network layers. This iterative process involves destroying the semantics of the input image with Gaussian noise. It also includes a denoising sampling process to return to a less noisy version of the input. This denoising sampling process forms the mathematical backbone of this class of models. Unlike static feedforward propagation, this approach allows for modeling the entire reverse denoising process and allows for capturing complex dependencies in data. This enables more expressive representations and enhanced modeling capabilities for further downstream tasks, such as classification.

For this reason, we explore the utility of diffusion models as representation learners for detecting deepfakes. We first, verify the performance of diffusion reconstructions for the purpose of detecting deepfakes. We also compare the results between this method and our latent noise-injected ResNet approach, alongside evaluating the usefulness of frequency transformations for this task.

2 Related Work

One of the earliest and most common approaches to detect generated images primarily focused on identifying image artifacts. These were mostly carried out using approaches that heavily rely on handcrafted feature design. Works such as [2,12,18,29], focused on this problem as a classification task. Although this worked, it failed to generalize properly to data other than the ones the classifier was trained on. Further research in this area led to better results such as the work by [33]. In this work, Pro-GAN [14] generated images were used to train multiple classifiers for each class of the LSUN bedroom dataset. Although this work performed well on GAN models and generalized to other GAN generated images the process was expensive.

This was followed by the advent of methods that analyzed the frequency domain of images such as the work proposed by [5,39]. This particular class of approaches relied on detecting artifacts present in the frequency domain of images, caused by up-sampling operations in GAN models. The authors of [19,38] proposed a method for detecting GAN-generated fingerprints in synthesized images. The ineffectiveness of this approach was then highlighted by [26] when it was used to attempt to detect images generated by the SoTA diffusion models. Although these techniques and even later works involving ensemble based approaches [17] generalize to some extent on GAN models, their performance on diffusion models is suboptimal.

Recent works, such as [34] and [8], however, compare the diffusion reconstruction error and use it either as a feature map to train a binary classifier or as a score threshold respectively to classify images from different distributions. In [34], the authors specifically suggest that, unlike autoregressive models, diffusion models learn the data distribution through a surrogate that is noise-based (by denoising at distinct time steps). In this work, we investigate the feasibility of using diffusion models to learn the distribution of data on a dataset generated with and without diffusion. In addition, we explore alternative methods for integrating latent spaces and noisy representations obtained by temporal inversion or perturbation. The aim of these approaches is to improve deepfake detection.

3 Method

We employ distinct frameworks to investigate the impact of incorporating a representation derived from a diffusion model. Initially, we examine the reconstruction of input images using the diffusion model before feeding them into a classification model. Subsequently, we introduce a classifier architecture with latent noise injected at various layers as a temporal perturbation regularizer. This is done to assess whether our model can capture potential trends in the noise generated by a diffusion model for a given image. Finally, we contrast these approaches with the utilization of a frequency transformation of the original images or the reconstruction errors.

Denoising diffusion probabilistic models (DDPMs) [11] have been successful at generating high-quality images without relying on adversarial learning. However, their drawback is the time-consuming simulation of a Markov chain to produce a sample. Diffusion implicit models (DDIMs) [28] have been introduced to speed up the sampling process. DDIMs are a more efficient class of iterative implicit probabilistic models that share the same learning procedure as DDPMs. Unlike DDPMs, DDIMs can perform semantically meaningful image interpolation by manipulating the initial latent variable. Overall, DDIMs offer a faster and more controlled way of generating high-quality images with meaningful interpolations.

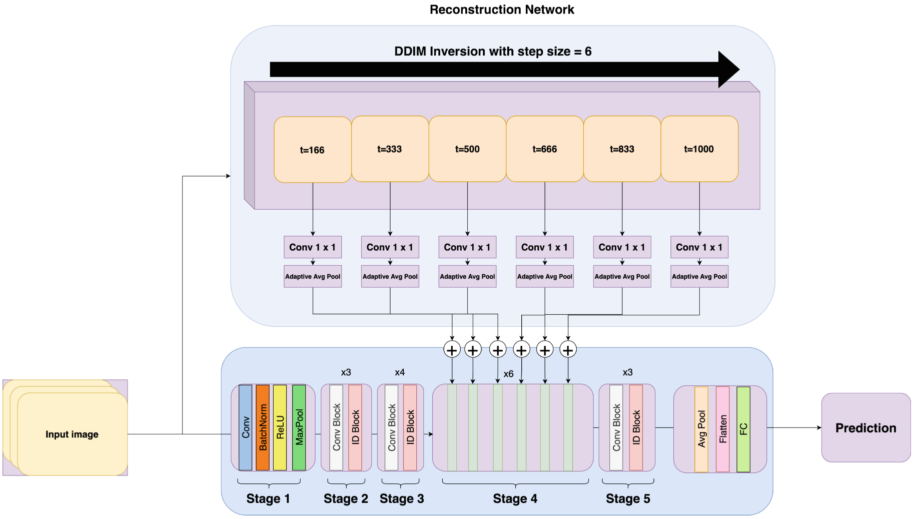


Fig. 1. ResNet50 [7] model with latent noise injected into first convolution layer of each of the 6 blocks in stage 4 of ResNet50. Our modified architecture for each block in stage 4 that adds latent noise to the input post-activation of the previous block.

More recently, inversion processes have gained attention for enhanced image reconstruction and editing [23]. The application of DDIM inversion is pivotal for extracting latent space representations in the diffusion domain, providing a deterministic approach. DDIM inversion involves executing DDIM sampling in reverse order, offering a valuable tool for improved image manipulation and reconstruction.

3.1 DDIM inversion

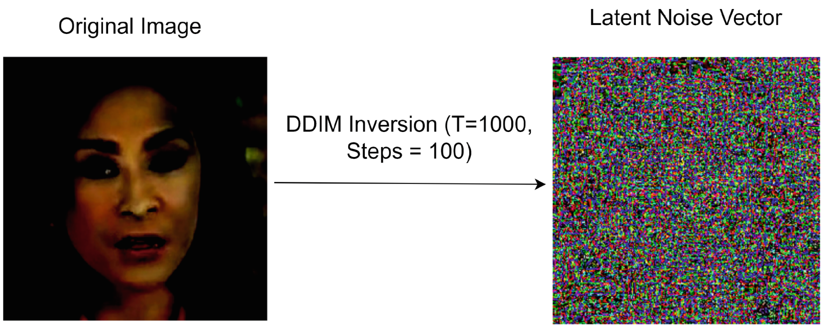


Fig. 2. DDIM inversion on a single normalized image from FF++ [27] to obtain it's latent space noise vector

A diffusion model comprises of a forward diffusion process (noising process) which degrades the input x_0 over timesteps $T (T = 1000)$ resulting in pure noise x_T and an associated reverse diffusion process (denoising process) that is used to generate back x_0 from x_T. Typically, the diffusion noising process is a closed form process given by equation 1. Here, x_t is the noisy version of our input x_0 at a specific timestep t, $\alpha_t = 1 - \beta_t$ where β_t denotes the forward noising schedule (linear, cosine etc) and therefore $\bar{\alpha}_t = \prod_{i=1}^{t} \alpha_i$ (Fig. 2).

$$x_t = \sqrt{\bar{\alpha}_t}\mathbf{x}_0 + \sqrt{1 - \bar{\alpha}_t}\boldsymbol{\epsilon} \tag{1}$$

The generalized sampling equation for diffusion models as described by the authors in [28] is given by equation 2 where σ_t controls the stochasticity of the sampling process and $\epsilon_\theta^{(t)}(x_t)$ is the model which takes noisy image x_t as input and predicts ϵ_t, the added noise.

$$\boldsymbol{x}_{t-1} = \sqrt{\alpha_{t-1}} \left(\frac{\boldsymbol{x}_t - \sqrt{1 - \alpha_t}\epsilon_\theta^{(t)}(\boldsymbol{x}_t)}{\sqrt{\alpha_t}} \right) + \sqrt{1 - \alpha_{t-1} - \sigma_t^2} \cdot \epsilon_\theta^{(t)}(\boldsymbol{x}_t) + \sigma_t \epsilon_t \tag{2}$$

If we substitute the value of σ_t as 0 (DDIM) in equation 2 we can configure a purely deterministic diffusion sampler given by equation 3, implying that for a unique noise vector the generated image will be unique as well.

$$\boldsymbol{x}_{t-1} = \sqrt{\frac{\alpha_{t-1}}{\alpha_t}}\boldsymbol{x}_t + \sqrt{\alpha_{t-1}} \left(\sqrt{\frac{1}{\alpha_{t-1}} - 1} - \sqrt{\frac{1}{\alpha_t} - 1} \right) \boldsymbol{\epsilon}_\theta(\boldsymbol{x}_t, t) \tag{3}$$

However, in DDIM inversion [21], we substitute the forward diffusion (noising) process described in equation 1 as shown below:

$$\boldsymbol{x}_{t+1} = \sqrt{\frac{\alpha_{t+1}}{\alpha_t}}\boldsymbol{x}_t + \sqrt{\alpha_{t+1}} \left(\sqrt{\frac{1}{\alpha_{t+1}} - 1} - \sqrt{\frac{1}{\alpha_t} - 1} \right) \boldsymbol{\epsilon}_\theta(\boldsymbol{x}_t, t) \tag{4}$$

As this suggests, now instead of performing a close form noising step for an image as shown in equation 1, we perform a model forward pass using the above sampling equation to obtain a noisy version of the image given at timestep t since the new form is reparameterized over ϵ_θ as opposed to $\epsilon \sim \mathcal{N}(0, I)$ in equation 1. Since this process is deterministic the reverse process adheres to the same data distribution as that of the input.

3.2 Reconstruction Network

Our classification model comprised of a learned reconstruction network followed by a ResNet50 [9] classifier with pretrained ImageNetV2 [24] weights. As illustrated in [28], due to the removal of the Markovian dependency for sampling operations, a subset of the total number of timesteps T any diffusion model was trained for, can be used for the generative (denoising/sampling) task. For

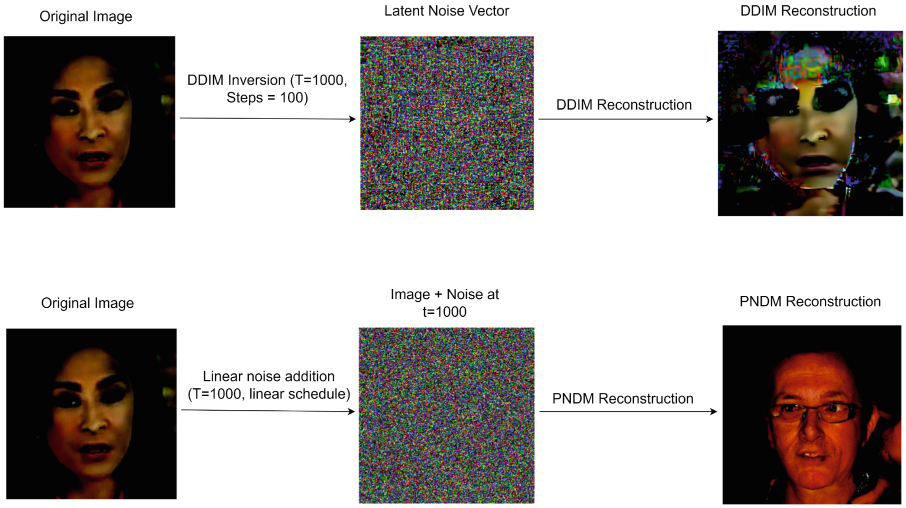


Fig. 3. Comparison of learned reconstruction using DDIM with deterministic denoising allowing us to sample the latent space for the image and therefore denoising it enables us to get back to the original image (top) vs. naive reconstruction using PNDM [16] where Gaussian noise is added and then denoised resulting in a completely new image as the process is not determinsitic without the use of DDIM (bottom)

$T = [1, 2, 3, ..., 1000]$ we can instead only sample on $\tau = [10, 20, 30, ..., 1000]$ for little or no loss in quality provided that the number of steps is suited to the complexity of the data. For our method, in the first step, we apply DDIM inversion to the input image with number of denoising steps represented as S. This step involves deriving the latent space representations of the input images. To generate the latent space noise vector, the model undergoes S forward passes in a non-Markovian, deterministic inverted denoising procedure. For the first step of our experiments involving PNDM we instead add Gaussian noise for 200 timesteps to obtain a noisy image. Subsequently, the obtained latent space or noisy image for DDIM and PNDM respectively is employed as the input for the pretrained denoising diffusion model and denoised using the sampler DDIM and PNDM [16] respectively, with a fixed step size, aiming to reconstruct the original images (Figure 3). This process completes the reconstruction by leveraging the information present in the latent noise vector or noisy image.

3.3 Latent ResNet

In addition to using reconstructed original images and their corresponding reconstruction errors, we also proposed using the latent space noise during the classifier training process. This idea is inspired by the original ResNet paper [9], which employed residual connections to facilitate the mapping of an identity function in deeper networks.

To implement this, we experiment with a “residual connection” mapping of the latent space noise vector to the convolution layers of the fourth stage in ResNet50. ResNet50 comprises of 5 distinct stages, each consisting of *Conv* and *Identity*(ID) blocks repeated a specific number of times. Stage 4 in particular comprises of 6 such blocks and we add our latent noise representations to the first *Conv* layer of each such block. Since there are 6 blocks we choose to invert the image across 6 timesteps since DDIM inversion allows us to sample a subset of the total number of steps. We choose ResNet50 since it is efficient and comparatively faster to train, which compensates for the time consuming diffusion process. The fourth stage of ResNet50 consists of 6 blocks as shown in Figure 1 as well as the fifth stage of ResNet50 that comprises of 3 blocks in a separate experiment. We refer to this in short as “Latent S4” and “Latent S5” respectively in our results table. It is known that the upper layers of convolution networks learn low level features while the higher level features are learned deeper into the network. Our assumption is that subtle changes in high level features are key in distinguishing between deepfakes and real images. Therefore, we add the latent space noise to the first convolution layer of each block in the fourth stage of ResNet50. We also make the choice of adding the latent noise vectors after the ReLU activation of the previous layer to ensure negative values do not get reduced to zero. The latent representation of each input image is obtained using DDIM inversion, a process that is repeated in six steps until a specified iteration limit ($T = 1000$) using a pretrained diffusion model. To ensure compatibility with the convolution layers, “1x1” convolutions are applied to the latent noise vectors, followed by a 2D adaptive average pooling layer. The denoised output from each of the six steps in the DDIM inversion process is then combined with the subsequent input for the convolution layers within the convolution and identity blocks of the fourth stage of the ResNet50 classifier. We also further experiment with a similar modification but for stage 5 of ResNet50 which we refer to as “Latent S5” in our experiments as mentioned above.

The frequency component of images is also an interesting avenue to explore for deepfake analysis, as demonstrated in works such as [5]. Hence, we also investigated high-frequency feature extraction by utilizing discrete Cosine and Fourier transforms on our original samples as well as our reconstruction error feature maps.

3.4 Frequency transforms

Typically the process of performing a 2D DFT on an image yields a symmetric result with the low frequencies at the corners of the image. This is due to the periodicity of the function itself. A traditional 1D DFT of a sequence of length $2n$ for example would produce a symmetric sequence of coefficients that comprise of real numbers, their complex and conjugate components. In 1D, the amplitude spectrum would generally be the moduli of the first $n+1$ coefficients. In contrast to this a 2D DFT is usually implemented in a separable fashion across specific dimensions (here we choose the spatial height and width dimensions). Each column of the resultant spectrum is akin to the 1D spectrum described above. Since images have many low-frequency features (high-level features) they tend to get concentrated in the corners. On centering, owing to the periodicity of the output changing from a range of $[0, 2\pi]$ to $[-\pi, \pi]$ the low-frequency components move from the corners to the center and manifest as a cross. Compared to the DFT, the DCT is instead used for compression. In essence, the higher level of detail spatial regions (e.g., nearly uniform blue sky in the photo of a sunny day) in an image are compressed while the low level features, spatial areas with high level of detail are exploded. To express this mathematically, a low-rank matrix is now used to represent the approximation of the same image matrix as before.

4 Experimentation and Results

The FF++ dataset [27] comprises of 1000 authentic video sequences. FF++ includes four manipulation methods to alter an existing authentic frame. The methods used to this end include Deepfakes [3], Face2Face [15], FaceSwap [31] and NeuralTextures [30]. This presents a uniquely different challenge for experimenting with our detection methods. It also allows us to measure and conclude if the current approaches of using reconstruction error as a means of distinguishing the sampling distribution from the data distribution is effective under a different set of circumstances as ones experimented on in [34]. Due to the longer inference time of diffusion models, we use a subset of the total frames present in the full FF++ dataset after normalizing their pixel values. We select, $90,126$ images for training, with $74,295$ images being deepfakes while $15,831$ are real. Our validation set comprises of $30,042$ images with $24,655$ and $5,387$ deepfakes and real images respectively. Finally our unseen test data comprises of $4,476$ images with 764 real images and $3,712$ deepfakes. The FF++ dataset we use for our experiments is clearly not balanced however, the class proportions are stratified across each split. In the literature since ROC AUC is the most widely used metric for FF++ dataset and because the class distribution roughly meets the requirement for using accuracy, we choose these as our metrics for evaluation. Furthermore, since the preprocessing step for many of these techniques involve capturing the face structure using the first frame of a video, there are also the presence of temporal artifacts in each sequence. Some of these temporal artifacts such as blurring and chromatic aberration also exist in the source video frames as well as the deepfakes with varying degrees. Furthermore, the roughly unbalanced nature of the dataset make it a challenging dataset for our task.

For our classifier we utilize a ResNet50 pretrained on ImageNetV2. Our input data is normalized for all our experiments. Furthermore, a batch size of 64 was used to finetune the classifier for 30 epochs in each experiment for parity. Binary cross entropy loss was used in conjunction with an L2 regularizer of $1\mathrm{e}{-4}$ for mixed precision training of the model.

We select a pretrained diffusion model, that has been trained on CELEBA-HQ (256×256) as the backbone of our reconstruction network for two of our experiments. We also perform further experiments by choosing the same backbone pretrained model for fine tuning since CELEBA-HQ contains high-quality facial images of celebrities captured in diverse lighting conditions, poses, and expressions. As such features like contours, textures are highly transferable to images in FF++ images due to the inherent similarity in the domain encompassing facial features. We finetuned this diffusion model on $12,403$ real images from the FF++ dataset with a batch size of 32 over 100 epochs. The noising schedule used was linear with minimum and maximum values of β_t as 0.0001 and 0.02 respectively. We used the AdamW optimizer with a learning rate of 1e−5 with an exponential learning rate with a gamma of 0.9. The maximum timestep for noising process was set to $T = 1000$. On input our finetuned diffusion model carried out DDIM inversion with a step size of $S = 10$ using the inverted DDIM sampler to first obtain the latent space representations of the input images. This implies, to generate our latent space noise vector we perform 10 model forward passes in a non markovian, deterministic inverted denoising process. Following this the non inverted DDIM sampler is used to reconstruct the original image from the latent noise vector. We used this model to test the efficacy of not only the reconstruction error but also the reconstructions obtained through the diffusion model. Finally, we experimented with our latent ResNet (S4 and S5) models to determine if temporal perturbations when training the model can help the model learn some discriminating features from the latent space inversions of the input data.

For our comparison with frequency transformed data, we used both Discrete Cosine Transform as well as Discrete Fourier Transform to augment our data before passing them into the same classifier as used for our other experiments. We apply DCT and DFT to both the normalized RGB input images as well as experiment with the DCT and DFT of the reconstruction error obtained between the normalized RGB images and their reconstructed versions. This is done to determine if the frequency domain has sufficient distinct features compared to the spatial domain to improve our classification task.

Upon experimentation, our stage 4 Latent ResNet model (Latent S4) incorporating latent space noise vectors generated through a diffusion model, pretrained on CELEBA-HQ (256) without further fine tuning, followed by adding their average to convolutional layers yielded the best results with an AUC of 0.9655. Our baseline model which includes ResNet50 classifier pretrained on ImageNetV2 performed the second best AUC score of 0.9606. The latent S4 model finetuned on FF++ obtained a slightly less good AUC of 0.9602, which leads us to conclude that finetuning was overall detrimental to the prior representations already learned by our pretrained diffusion backbone. The Latent S4 model also outperformed our Latent S5 model. The diffusion reconstruction error as proposed in [34] performed suboptimally compared to our baseline and obtained an AUC of 0.9085. Evidently, on this dataset the DFT and DCT of the images were also outperformed, as frequency domain features between the deepfakes and the real images were much harder features to learn from compared to spatial inconsistencies. We also observed that utilizing similar frequency transforms on the reconstruction error did not improve performance. An interesting point to note, is that compared to the study conducted on diffusion generated LSUN bedroom

Table 1. Classification performance comparison on FF++ dataset using learned DDIM reconstructions.

Method	Accuracy	AUC
Baseline	92.25	0.9606
DDIM Reconstruction	82.93	0.5968
DIRE [34]	89.10	0.9085
DFT	82.93	0.6675
DCT	83.45	0.7197
DFT + DIRE	82.93	0.5888
DCT + DIRE	82.93	0.5948
Latent S4	**93.17**	**0.9655**
Latent S5	91.95	0.9483
Latent S4 (FF++ finetuned)	92.47	0.9602
Latent S5 (FF++ finetuned)	92.87	0.9604

[37] deepfakes in [34], the performance of diffusion reconstruction error deteriorated compared to our baseline for FF++. Moreover, we observed that with the reconstructions themselves having a lower AUC of 0.5968, finding an absolute difference between them and the original images only makes the performance of the model decrease compared to our baseline.

Table 2. Classification performance comparison on LSUNB dataset using naive PNDM reconstructions.

Method	Accuracy	AUC
Baseline	100	1.0
PNDM Reconstruction	93.44	0.9876
Naive DIRE	96.56	0.9940
DFT	99.97	1.0
DCT	100	1.0
DFT + Naive DIRE	80.11	0.9028
DCT + Naive DIRE	82.32	0.8886

Thus the diffusion reconstructions only impede our models performance. Further experimentation was also carried out using naive fourth order PNDM [16] reconstructions (as shown in table 2) on diffusion generated LSUN bedroom images. We sampled $30,000$ real images from LSUN-B alongside $30,000$ fake/synthetic images using an unconditional diffusion model using DDIM for faster generation. For this experiment we used the U-Net from the same pre-trained diffusion model that had been trained on LSUN-B images without further finetuning as the backbone for our reconstruction network. Although for generation of our fake data we used DDIM sampling, we utilize a naive reconstruction approach using PNDM for our reconstruction network. For this purpose we add Gaussian noise to our input images for 200 steps using the traditional [11] process with a linear schedule followed by a strided PNDM denoising over $S = 20$ steps. In both circumstances, although the diffusion reconstruction led to a high AUC overall, it performed worse than the baseline of just using the original images, signifying that the reconstructions only impeded performance for our classifier. Despite the poor performance of reconstruction error, we did observe that the latent ResNet architecture performed almost at parity with our baseline. This indicates that the latent space noise vectors could be a better representation compared to the reconstructions which might induce more variance in the data. Furthermore, although diffusion reconstruction error performed better than traditional frequency based approaches like DCT and DFT, the reconstructions by themselves exhibit unfavorable performance.

5 Conclusion

In this paper, we showcased, the performance of diffusion model based representations for a discriminative task. We observe that despite good performance by diffusion reconstruction error, an approach primarily highlighted in [34] for diffusion based deepfakes, their performance cannot be attributed to the diffusion reconstructions. Instead, they only benefit from spatially classifiable features. Traditional methods like DCT and DFT achieved higher scores than reconstructions, plausibly because frequency domain approaches preserve high frequency inconsistencies in the data to a higher degree than diffusion reconstructions. Diffusion models have an inductive bias that allows them to preserve low-frequency features first and then recover high-frequency details. This results in a biased recovery of frequencies that are in a minority as explained in [36]. However, using latent vectors instead as a regularizer during classification resulted in competitive performance to our baseline.

As a future prospective, we intend to experiment with frequency components of the latent vectors instead of diffusion reconstructions since they fundamentally introduce more bias in the sampling process due to more model forward passes. Another aspect of the current pipeline we want to improve is latency, post training quantization to $int8$ [32] for convolutional layers as well as linear layers alongside layer fusing can lead to faster inference times and newer state-of-the-art approaches to quantizing diffusion models, post training can also be explored [10]. Furthermore, we plan to make use of better, more comprehensive datasets that curate both high quality manipulated images and high quality synthetic images in order to test the robustness of the proposed approach.

Acknowledgment. We gratefully acknowledge the support of the Computer Research Institute of Montreal (CRIM), the Ministère de l'Économie, de l'Innovation et de l'Énergie (MEIE) of Quebec, and the Natural Sciences and Engineering Research Council of Canada (NSERC).

References

1. Afchar, D., Nozick, V., Yamagishi, J., Echizen, I.: Mesonet: a compact facial video forgery detection network. In: 2018 IEEE international workshop on information forensics and security (WIFS). pp. 1–7. IEEE (2018)
2. Chollet, F.: Xception: Deep learning with depthwise separable convolutions. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 1800–1807. IEEE Computer Society, Los Alamitos, CA, USA (jul 2017). https://doi.org/10.1109/CVPR.2017.195, https://doi.ieeecomputersociety.org/10.1109/CVPR.2017.195
3. Deepfakes: Faceswap: Deepfake your own images and videos. https://github.com/deepfakes/faceswap (2023)
4. Dhariwal, P., Nichol, A.Q.: Diffusion models beat GANs on image synthesis. In: Beygelzimer, A., Dauphin, Y., Liang, P., Vaughan, J.W. (eds.) Advances in Neural Information Processing Systems (2021), https://openreview.net/forum?id=AAWuCvzaVt
5. Frank, J., Eisenhofer, T., Schönherr, L., Fischer, A., Kolossa, D., Holz, T.: Leveraging frequency analysis for deep fake image recognition. In: III, H.D., Singh, A. (eds.) Proceedings of the 37th International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 119, pp. 3247–3258. PMLR (13–18 Jul 2020), https://proceedings.mlr.press/v119/frank20a.html
6. Giudice, O., Guarnera, L., Battiato, S.: Fighting deepfakes by detecting gan dct anomalies. Journal of Imaging **7**(8) (2021). https://doi.org/10.3390/jimaging7080128, https://www.mdpi.com/2313-433X/7/8/128
7. Gomes, R., Kamrowski, C., Langlois, J., Rozario, P., Dircks, I., Grottodden, K., Martinez, M., Tee, W.Z., Sargeant, K., LaFleur, C., Haley, M.: A comprehensive review of machine learning used to combat covid-19. Diagnostics **12**(8) (2022). https://doi.org/10.3390/diagnostics12081853, https://www.mdpi.com/2075-4418/12/8/1853
8. Graham, M.S., Pinaya, W.H., Tudosiu, P.D., Nachev, P., Ourselin, S., Cardoso, J.: Denoising diffusion models for out-of-distribution detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops. pp. 2947–2956 (June 2023)
9. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 770–778 (2016). https://doi.org/10.1109/CVPR.2016.90
10. He, Y., Liu, L., Liu, J., Wu, W., Zhou, H., Zhuang, B.: PTQD: Accurate post-training quantization for diffusion models. In: Oh, A., Naumann, T., Globerson, A., Saenko, K., Hardt, M., Levine, S. (eds.) Advances in Neural Information Processing Systems. vol. 36, pp. 13237–13249. Curran Associates, Inc. (2023)
11. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. In: Proceedings of the 34th International Conference on Neural Information Processing Systems. NIPS '20, Curran Associates Inc., Red Hook, NY, USA (2020)
12. Huang, G., Liu, Z., van der Maaten, L., Weinberger, K.Q.: Densely connected convolutional networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2017)
13. Iqbal, F., Abbasi, A., Javed, A.R., Almadhor, A., Jalil, Z., Anwar, S., Rida, I.: Data augmentation-based novel deep learning method for deepfaked images detection. ACM Trans. Multimedia Comput. Commun. Appl. (apr 2023). https://doi.org/10.1145/3592615, https://doi.org/10.1145/3592615, just Accepted

14. Karras, T., Aila, T., Laine, S., Lehtinen, J.: Progressive growing of GANs for improved quality, stability, and variation. In: International Conference on Learning Representations (2018), https://openreview.net/forum?id=Hk99zCeAb
15. Kowalski, M.: Faceswap: Face swapping and image manipulation tool. https://github.com/MarekKowalski/FaceSwap (2023)
16. Liu, L., Ren, Y., Lin, Z., Zhao, Z.: Pseudo numerical methods for diffusion models on manifolds. In: International Conference on Learning Representations (2022)
17. Mandelli, S., Bonettini, N., Bestagini, P., Tubaro, S.: Detecting gan-generated images by orthogonal training of multiple cnns. In: 2022 IEEE International Conference on Image Processing (ICIP). pp. 3091–3095 (2022). https://doi.org/10.1109/ICIP46576.2022.9897310
18. Marra, F., Gragnaniello, D., Cozzolino, D., Verdoliva, L.: Detection of gan-generated fake images over social networks. In: 2018 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR). pp. 384–389 (2018). https://doi.org/10.1109/MIPR.2018.00084
19. Marra, F., Gragnaniello, D., Verdoliva, L., Poggi, G.: Do gans leave artificial fingerprints? 2019 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR) pp. 506–511 (2018), https://api.semanticscholar.org/CorpusID:57189570
20. Meng, C., He, Y., Song, Y., Song, J., Wu, J., Zhu, J.Y., Ermon, S.: SDEdit: Guided image synthesis and editing with stochastic differential equations. In: International Conference on Learning Representations (2022), https://openreview.net/forum?id=aBsCjcPu_tE
21. Miyake, D., Iohara, A., Saito, Y., Tanaka, T.: Negative-prompt inversion: Fast image inversion for editing with text-guided diffusion models (2023), https://arxiv.org/abs/2305.16807
22. Nirkin, Y., Keller, Y., Hassner, T.: FSGAN: Subject agnostic face swapping and reenactment. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 7184–7193 (2019)
23. Parmar, G., Singh, K.K., Zhang, R., Li, Y., Lu, J., Zhu, J.Y.: Zero-shot image-to-image translation (2023), https://arxiv.org/abs/2302.03027
24. Recht, B., Roelofs, R., Schmidt, L., Shankar, V.: Do ImageNet classifiers generalize to ImageNet? In: Chaudhuri, K., Salakhutdinov, R. (eds.) Proceedings of the 36th International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 97, pp. 5389–5400. PMLR (09–15 Jun 2019), https://proceedings.mlr.press/v97/recht19a.html
25. Ricker, J., Damm, S., Holz, T., Fischer, A.: Towards the detection of diffusion model deepfakes. ArXiv **abs/2210.14571** (2022), https://api.semanticscholar.org/CorpusID:253116680
26. Ricker, J., Damm, S., Holz, T., Fischer, A.: Towards the detection of diffusion model deepfakes (2023)
27. Rossler, A., Cozzolino, D., Verdoliva, L., Riess, C., Thies, J., Nießner, M.: Faceforensics++: Learning to detect manipulated facial images. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 1–11 (2019)
28. Song, J., Meng, C., Ermon, S.: Denoising diffusion implicit models. In: International Conference on Learning Representations (2021), https://openreview.net/forum?id=St1giarCHLP
29. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z.: Rethinking the inception architecture for computer vision. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 2818–2826 (2016). https://doi.org/10.1109/CVPR.2016.308

30. Thies, J., Zollhöfer, M., Nießner, M.: Deferred neural rendering: image synthesis using neural textures. ACM Trans. Graph. **38**(4) (jul 2019). https://doi.org/10.1145/3306346.3323035, https://doi.org/10.1145/3306346.3323035
31. Thies, J., Zollhofer, M., Stamminger, M., Theobalt, C., Nießner, M.: Face2face: Real-time face capture and reenactment of rgb videos. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2387–2395 (2016)
32. Van Baalen, M., Kuzmin, A., Nair, S.S., Ren, Y., Mahurin, E., Patel, C., Subramanian, S., Lee, S., Nagel, M., Soriaga, J., Blankevoort, T.: Fp8 versus int8 for efficient deep learning inference (2023), https://arxiv.org/abs/2303.17951
33. Wang, S.Y., Wang, O., Zhang, R., Owens, A., Efros, A.A.: Cnn-generated images are surprisingly easy to spot...for now. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 8692–8701 (2020). https://doi.org/10.1109/CVPR42600.2020.00872
34. Wang, Z., Bao, J., gang Zhou, W., Wang, W., Hu, H., Chen, H., Li, H.: Dire for diffusion-generated image detection. 2023 IEEE/CVF International Conference on Computer Vision (ICCV) pp. 22388–22398 (2023), https://api.semanticscholar.org/CorpusID:257557819
35. Wesselkamp, V., Rieck, K., Arp, D., Quiring, E.: Misleading deep-fake detection with gan fingerprints. In: 2022 IEEE Security and Privacy Workshops (SPW). pp. 59–65. IEEE Computer Society, Los Alamitos, CA, USA (may 2022). https://doi.org/10.1109/SPW54247.2022.9833860
36. Yang, X., Zhou, D., Feng, J., Wang, X.: Diffusion probabilistic model made slim. In: 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 22552–22562. IEEE Computer Society, Los Alamitos, CA, USA (jun 2023). https://doi.org/10.1109/CVPR52729.2023.02160, https://doi.ieeecomputersociety.org/10.1109/CVPR52729.2023.02160
37. Yu, F., Zhang, Y., Song, S., Seff, A., Xiao, J.: LSUN: construction of a large-scale image dataset using deep learning with humans in the loop. CoRR **abs/1506.03365** (2015), http://arxiv.org/abs/1506.03365
38. Yu, N., Davis, L., Fritz, M.: Attributing fake images to gans: Learning and analyzing gan fingerprints. In: 2019 IEEE/CVF International Conference on Computer Vision (ICCV). pp. 7555–7565. IEEE Computer Society, Los Alamitos, CA, USA (nov 2019). https://doi.org/10.1109/ICCV.2019.00765, https://doi.ieeecomputersociety.org/10.1109/ICCV.2019.00765
39. Zhang, X., Karaman, S., Chang, S.F.: Detecting and simulating artifacts in gan fake images. In: 2019 IEEE International Workshop on Information Forensics and Security (WIFS). pp. 1–6 (2019). https://doi.org/10.1109/WIFS47025.2019.9035107

Hybrid Transformer-CNN-Based Attention in Video Turbulence Mitigation (HATM)

Mohammad Ahangar Kiasari[1], Khan Muhammad[3](✉), Sambit Bakshi[4], and Ik Hyun Lee[1,2](✉)

[1] Department of Mechatronics Engineering, Tech University of Korea, Siheung 15073, South Korea
{ahangar100,ihlee}@tukorea.ac.kr
[2] IKLAB Inc., Seoul 08513, South Korea
[3] School of Convergence, Sungkyunkwan University, Seoul 03063, South Korea
khanmuhammad@g.skku.edu
[4] Department of Computer Science and Engineering, National Institute of Technology Rourkela, Rourkela, India
bakshisambit@nitrkl.ac.in

Abstract. This study introduces a hybrid deep learning framework for turbulence mitigation (HATM) in videos, integrating a transformer-based followed by CNN-based attention modules. Due to the computational demands associated with transformers, we propose a simple technique within the transformer module to enhance computational efficiency. Additionally, to better exploit spatial and channel information, we introduce a CNN-attention module which captures global and local inter- and intra-frame dependencies. The overall structure of the model follows U-net, while the skip connections are replaced by our attention blocks to further explore local, spatial, and temporal dependencies. Our model is trained on a simulated turbulence dataset and evaluated on both simulated and real-world datasets to gauge its generalization performance. The effectiveness of each component within our model is also evaluated through ablation studies. Experimental outputs show that our model improves PSNR and SSIM scores, and notably enhances the reconstruction of text images, making the restored text images more readable and cleaner. Overall, our HATM framework represents an advancement towards addressing turbulence distortion in video sequences, showcasing improvements both qualitatively and quantitatively, and offering promising solutions for various applications requiring enhanced video content restoration and mitigation of turbulence-induced artifacts.

Keywords: Video turbulence mitigation · Transformer · Hybrid Attention

1 Introduction

Video turbulence mitigation is the process of eliminating tilting, blurriness, noise, and other visual distortions caused by atmospheric turbulence particularly when capturing videos from long distances. Conventional image processing

A. Antonacopoulos et al. (Eds.): ICPR 2024, LNCS 15321, pp. 242–256, 2025.
https://doi.org/10.1007/978-3-031-78305-0_16

techniques [8,32,33] employed for turbulence elimination are constrained in their ability to overcome complex turbulence distortions. Consequently, researchers have turned their attention towards more advanced models e.g., Convolutional Neural Networks (CNNs) and Transformers.

Recently, Transformers have exhibited superior performance in capturing spatial and temporal features, thereby achieving high performance in addressing many applications such as text detection and segmentation [28,36], object localization [25], and classification [21], etc. In the realm of image and video enhancements, transformer models also demonstrated remarkable performances [5,29,35,38]. For instance, as a de-blurring model, S. W. Amir et al. presented a transformer based architecture "Restormer" to efficiently exploit correlations among long-distance features [35]. Chen et al. proposed "DAT" which includes consecutive Transformer blocks separately exploring spatial and channel correlations using the self-attention technique for image super-resolution [5]. Recently, Zhang et al. introduced TMT [37], a multi-stage deep learning (DL) method, involving de-warping and de-blurring architectures. In the de-warping model, they employed a simple U-net structure, while for the de-blurring, they integrated the Restormer [35] structure, introducing a shuffle attention mechanism to incorporate temporal correlations between frames.

Besides advancements in transformer-based models, researchers have also explored CNNs architectures [18,38]. For example, Zhong et al. presented a deep convolutional recurrent architecture for eliminating blurring of adjacent frames in videos [38]. Liu et al. [18] introduced fully convolutional networks incorporating multi-scale features fusion, which integrates information from low to high resolution and vice versa to effectively eliminate blurriness in video frames. Their results demonstrate superior performance compared to several state-of-the-art (SOTA) de-blurring algorithms. Son et al. [29] also proposed a novel video de-blurring network based on motion estimation compensation and showed the superiority of their model over other cutting-edge models in video de-blurring. For turbulence mitigation, Vint et al. initially applied existing DL approaches, originally designed for noise and blurriness removal tasks [31]. Mao et al. subsequently introduced a simplified U-net structure, which operates on 50 input frames and employs mean square error as the loss function during supervised training [19]. They also proposed the phase-to-space (P2S) transform for generating training datasets, a technique that we also used in our study.

This paper introduces a hybrid attention model for turbulence mitigation (HATM) by integrating a simplified attention module [4] and a transformer module [19,35] to leverage both structures for enhancing video turbulence mitigation. In terms of intra-block design, recognizing the substantial computational demands associated with computing transformer-based attention, we introduce a simple strategy to mitigate the computational complexity of the Transformer module while preserving the performance. Additionally, the integration of a simplified CNN-based attention mechanism within each block demonstrated an improvement in the final video quality. Regarding inter-block architecture, unlike [19], we replace skip connections with an attention block to perform a more in-depth analysis of low-level features. We conduct experiments by training and evaluating the proposed model using our simulated dataset. Additionally, we perform an ablation study to illustrate the effectiveness of the proposed attention block in achieving higher PSNR and SSIM scores. We also used another dataset (BVI-CLEAR [2]) to evaluate the generalization performance of the proposed model compared to recent SOTA models. The results show the superior visual performance of the proposed model, particularly in restoring turbulence text image sequences. The key highlights of HATM can be outlined as the following:

- Introducing an attention mechanism incorporating a transformer-based followed by a CNN-based attention modules.
- Introducing a simple technique within the transformer-based attention module for reducing computational load.
- Evaluating the proposed structure through experiments on three datasets including static and dynamic videos to showcase the impact of the proposed model on turbulence elimination and validate its generalization.

This paper unfolds in the following sections: First, Section 2 critically discusses the related works. Then, Section 3 introduces the methods and materials. In Section 4, detailed information about the dataset, implementation, and comparisons is presented. Finally, the last section delves into a discussion of the research achievements.

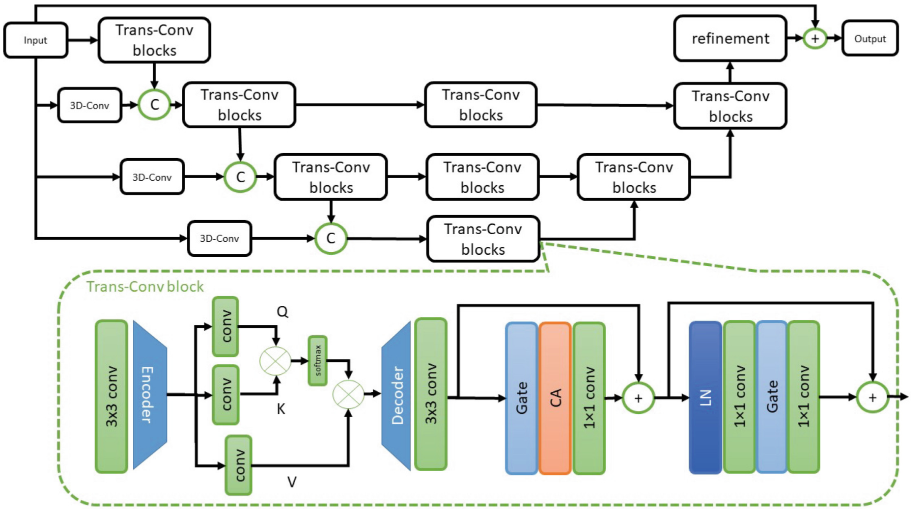


Fig. 1. Structure of the proposed Hybrid Transformer-Convolutional Turbulence Mitigation Model (HATM). CA and LN indicate channel attention and layer normalization, respectively.

2 Related works

Various attention mechanisms are extensively utilized in enhancing images and videos restorations. Recent advancements in deep enhancement techniques can be categorized into two main types: CNN-based and transformer-based models. This section provides a brief analysis of these DL models.

2.1 CNN-based Image and Video Enhancement

In image enhancement, convolutional structures such as U-net have been widely used in recent years. For example, Guo et al. [9] applied the U-net structure to restore high-light images from low-light ones, aiming to enhance object detection performance. They applied high-light enhancement techniques as a preprocessing stage for detection. Moran et al. [22] used convolutional structures to enhance image properties such as color and luminance. There is also recent image enhancement research introduced by Chen et al. [6], which applied CNN-based attention techniques for the image enhancement task. Wang and his colleagues also proposed a pioneering generative adversarial framework specifically designed for image restoration [34]. Their approach involved the removal of batch normalization from the residual units, coupled with the incorporation of perceptual loss before the ReLU activation functions. These strategies enable the model to further enhance the brightness and sharpness of the final outputs, leading to visually appealing results. Park et al. introduced an inner-recurrence module inside an recurrent neural networks (RNN) cell to cope with the constraints associated with short-term memory limitations [23]. This innovative approach enhances the ability of exploring large dependencies, thereby improving its performance on tasks requiring temporal coherence. Additionally, Liu et al. [18] presented a novel approach aimed at improving de-blurring in dynamic video frames. They introduced blur-invariant motion estimation and Pixel Volume modules to effectively utilize motion estimation, resulting in sharper frames with reduced artifacts. This innovative technique demonstrates significant advancements in dynamic video de-blurring, paving the way for improved video quality in various applications.

2.2 Transformer-based Image and Video Enhancement

Recently, video enhancement utilizing transformers has emerged as a cutting-edge technique in computer vision. Originally designed for natural language processing tasks, transformers have been repurposed to effectively model temporal dependencies in image and video sequences. Using self-attention, transformers can explore distant spatial and spectral dependencies and efficiently propagate information across frames [24]. For instance, Souibgui et al. [30] exploited the transformer's capability to enhance textual image reconstruction from distorted document images. In the realm of turbulence mitigation, Zhang et al. [37] introduced a hybrid transformer-CNN module embedded within a U-net structure to address turbulence distortions in sequential frames. They adopted the hybrid structure initially proposed in [35], which was originally applied for image restoration. Furthermore, they introduced a novel shuffle attention block to efficiently harness spatial and sequential information in video sequences during the training process. In our paper, we have enhanced both the transformer and CNN components of the model to further improve performance.

3 Methodology

Fig. 1 illustrates an overview of the proposed turbulence mitigation model (HATM). The model takes turbulence video frames as an input and produces the corresponding restored frames. In this section, we introduce transformer-based and convolution-based attention modules which are subsequently applied in a single attention block.

3.1 Transformer-based Attention Module

The structure of the transformer module is detailed as follows: Within each transformer module, the input feature map is denoted as $F \in \mathbb{R}^{T\times C\times H\times W}$, where T represents the frame number, C represents the channel number, and H × W represents the spatial dimensions of the feature map. The transformer module initiation involves two depth-wise 3D convolution layers with dilation factors 2 and 3, designed to augment the receptive field. Subsequently, a 3D depth-wise encoder layer is employed to reduce the spatial dimensions of the feature map by a factor of 1/2. This decoder is introduced to alleviate the computational load associated with computing the self-attention map. The output of the encoder is denoted as $F_{enc} \in \mathbb{R}^{T\times C\times H/2\times W/2}$.

According to the self-attention techniques [7,10,13,27], the computation of *Query* (Q), *Key* (K), and *Value* (V) involves reshaping the feature map to $(C\times T, W/2\times H/2)$. The individual components are calculated as $Q = W^q\times F_{enc}$, $K = W^k\times F_{enc}$, and $V = W^v\times F_{enc}$, where, $W^q \in \mathbb{R}^{CT\times HW/4}$, $W^k \in \mathbb{R}^{CT\times HW/4}$, and $W^v \in \mathbb{R}^{CT\times HW/4}$. The recalculation of X is performed utilizing the self-attention map in the following manner:

$$\hat{F}_{enc} = Softmax(\frac{QK^T}{\alpha})\times V, \tag{1}$$

where, α is a learnable scaling parameter to control the magnitude and $\hat{X}$ is the output of the self-attention operation. The $\hat{F}_{enc}$ is reshaped to $T\times C\times H/2\times W/2$ and then fed into the decoder layer to upsample $\hat{F}_{enc}$ to the original spatial size $H\times W$. The computational complexity of the above attention mechanism is expressed as $O(\frac{HW\times(CT)^2}{h})$. So, reducing spatial sizes to $H/2$ and $W/2$ leads to a quadratic decrease in computational complexity.

3.2 Convolution-based Attention

In the proposed framework, we introduce a CNN-based module designed to facilitate the interaction of global and channel-specific information within the feature space. To achieve this objective, we implement a simplified attention module [4], comprising two primary components: a simplified channel attention mechanism and a gated operation defined as $Gate(X, Y) = X \odot Y$. The computation of X and Y involves a 3D convolution that condenses spatial information into channels, followed by a division of features across channels. The resulting output from the $Gate()$ operator feeds into the CNN-based channel attention, where the feature map is compressed in channels and then multiplied with the original feature map. This facilitates the model in focusing more attention on significant channels. Note that, in Fig. 1, LN indicates layer normalization on 3D feature space.

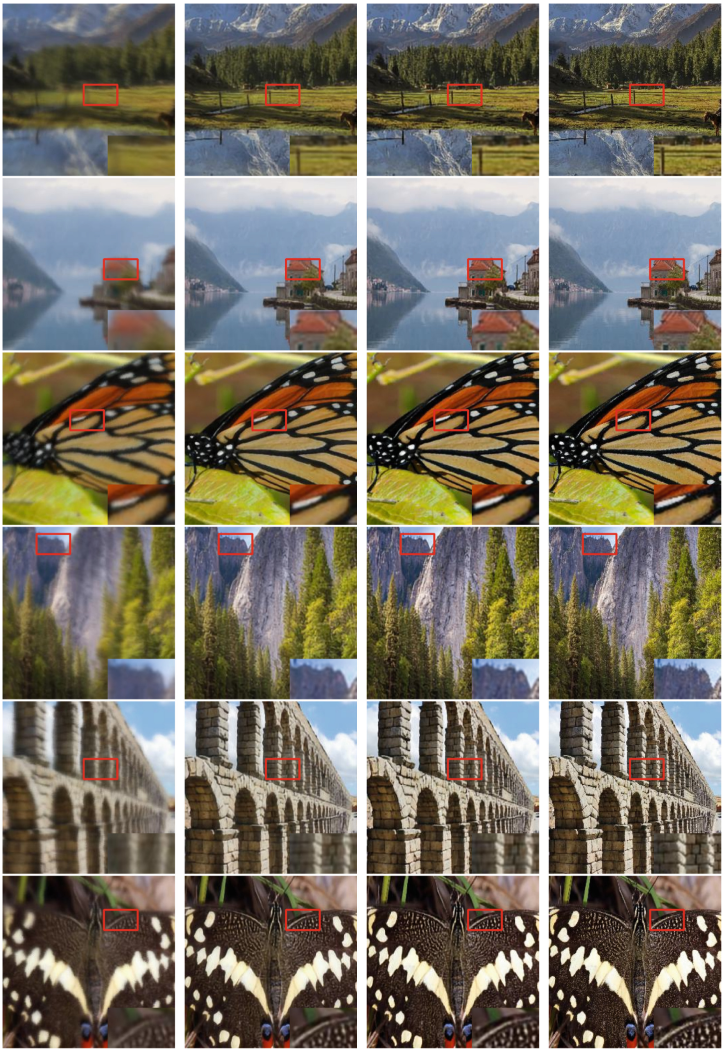
The overall structure of the proposed model follows U-net. Considering the significance of enhancing low-level features to improve PSNR scores, we replace the skip connections, which bridge encoder to decoder layers, with the proposed attention blocks. Moreover, in image and video enhancement, especially in turbulence mitigation and de-blurring algorithms, researchers commonly apply skip connections from the input to the output. Based on our experience, adding this skip connection can significantly speed up the training process and enhance performance. Consequently, following other recent models [2] in this area, we incorporate the skip connection in our structure. We also apply group convolution in all 3D convolution layers to address the complexity. This approach reduces computational complexity and memory usage by dividing input channels into smaller groups channels and performing convolutions separately. Consequently, processing times are faster, and fewer parameters are needed, increasing the applicability of our HATM.

4 Results and Discussion

In this section, we conduct experiments to evaluate the proposed model, as discussed earlier in previous section. We applied DIV2K dataset [1] to generate the test and training videos, which is explained in Section 4.1. For evaluating the proposed model, we apply three datasets including the BVI-CLEAR [3], the real world turbulence [2,11] dataset, and the real-world text dataset [20].

4.1 Data Preparation

For training purposes, our dataset was expanded by incorporating images from the DIV2K dataset [1]. De-blurring and de-noising degradation functions were applied to images using the provided TurbulenceSIMP2S simulator [19]. We collected images from the DIV2K dataset [1] and generated simulated turbulence video frames using the TurbulenceSIMP2S simulator [19]. This software offers degradation functions such as blurring, warping, and noise. Initially, 800



(a) Input (b) TMT [37] (c) HATM (ours) (d) GT

Fig. 2. Qualitative comparison between the proposed model (HATM) and TMT [37] using simulated static sequential images. (a) represents first frame of input sequential images. (b) shows the output of TMT [37]. (c) represents the output of the proposed model (HATM). (d) represents the corresponding GT images.

images were selected from the DIV2K dataset, and images are transformed into a sequence of distorted frames. We generate 50 disturbed frames per image, simulating blurring, noise, and warping effects. Additionally, different distortion levels were applied to further augment the dataset. Out of the 800 videos generated from each image, 20 videos were randomly selected for testing, and the remaining videos were allocated for training.

To generate dynamic turbulence videos, we gather 300 videos from the Sport1M dataset [12], a comprehensive action recognition dataset comprising 1.1 million videos spanning 487 sport action classes. Among them, 250 videos are selected. A partition of 200 videos is assigned for training purposes, while the remaining videos are designated for testing. Each video is resized to 512 × 512. Since each video has different length, first 100 frames of each video are selected for the training and testing. Similar to the static dataset, for further augmenting the dataset, various distortion levels and conventional augmentation techniques, including cropping, flipping, and noise addition, are applied during the training phase.

In addition to the simulated datasets, we included other turbulence video datasets in the test phase to assess the generalization capability of the proposed model. Specifically, we utilized the BVI-CLEAR [3] Dataset, which comprises eight distinct static turbulence videos, including *barcodes*, *books*, *boxes*, *carback*, *carfront*, *faces*, *plant*, and *toys*. We additionally apply another real-world dataset [2,11] comprising 14 turbulence videos, encompassing both static and dynamic scenarios. We present results obtained from the test datasets to visually demonstrate the generalization performance of the proposed model.

Note that, for the training phase of the static model, only synthetic static data is utilized. Since dynamic videos can include both static and dynamic scenes, the model, intended for turbulence removal in dynamic videos, is trained on a combination of synthetic dynamic and static datasets.

4.2 Comparison With Recent Models

We conduct qualitative and quantitative comparisons in Sections 4.2, 4.3, and 4.4 using some recent models such as Dnet [14], NDIR [15], VRT [16], U-net [26], RVRT [17], and TMT [37]. U-net [26], a widely used reference for enhancing image and video processing, RVRT [17], a recent video restoration model, and TMT [37], recognized as the current SOTA model for video turbulence mitigation. Dnet and NDIR are both developed for turbulence mitigation and VRT is originally designed for video restoration. It is important to note that we trained all models with our training dataset to ensure a fair comparison. Fig. 2 illustrates the performance of TMT and the proposed model on our synthesized dataset and Table 1 presents the average objective qualities, including PSNR and SSIM, of the restored and ground truth static videos. Note that, all models are trained with the dataset outlined in Section 4.1 and the number of input frames per video is set to 12 for all experiments conducted in this study. The results demonstrate that our HATM achieves higher PSNR and SSIM scores outperforming other U-net, TMT, and RVRT models on our synthesized datasets.

We also conduct turbulence mitigation experiments on dynamic datasets. To address turbulence removal in dynamic videos, we train the model using both static and dynamic synthetic datasets. The results in Table 2 demonstrate that our model outperforms TMT and has performance comparable to RVRT.

Table 1. Average inference PSNR and SSIM scores of different models across static video scenes with varying distortion levels. Input image size is 256×256 and the number of input frames is 12 for all experiments.

Methods	Turbulence level			
	D=0.1, r=0.1		D=0.1, r=0.2	
	PSNR	SSIM	PSNR	SSIM
U-net [26]	23.32	0.7254	23.37	0.7379
TMT [37]	23.64	0.7412	24.14	0.7519
RVRT [17]	23.57	0.7401	23.98	0.7448
HATM (ours)	24.03	0.7798	26.03	0.8521

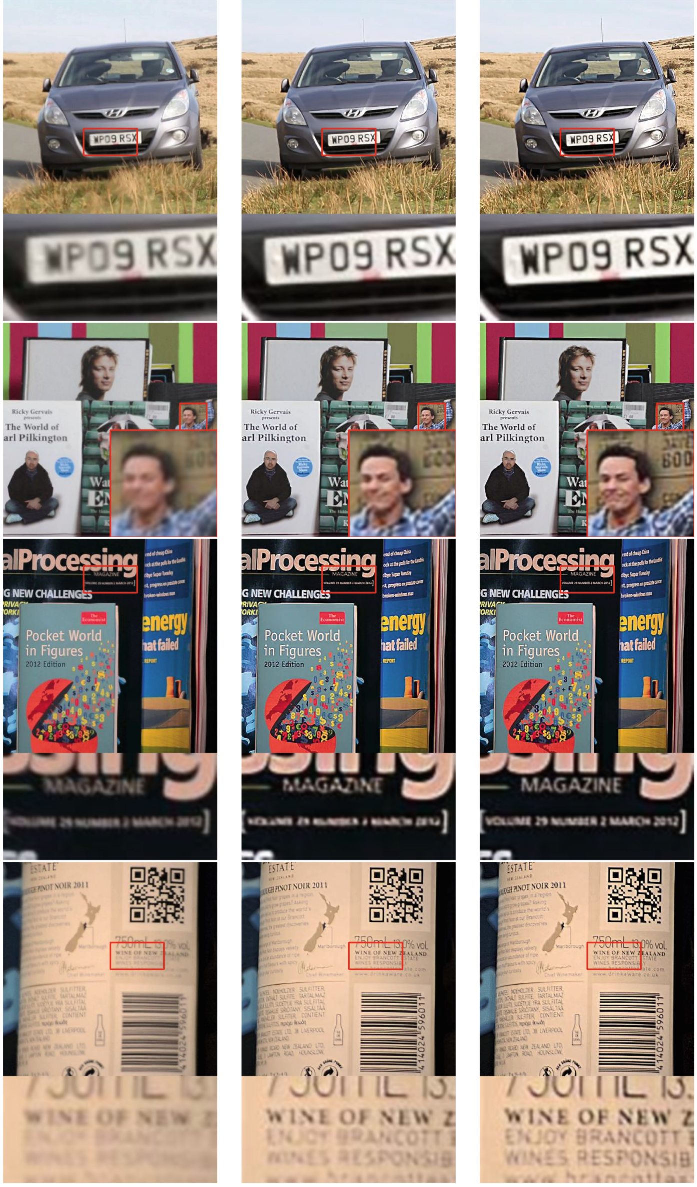
4.3 Ablation Study on Different Attention Modules

In our proposed model, we introduced a block attention mechanism comprised of two attention modules: CNN-based and Transformer-based attentions. This section explores the influence of this hybrid attention block on the PSNR and SSIM scores. Table 3 presents the outcomes of models employing three distinct attention blocks. The first model exclusively utilizes a CNN-based attention block, the second model follows the approach outlined in [37], incorporating a transformer block with a Gated Feed-forward (GF) [4,35], and the third model represents our proposed architecture which uses the transformer with the CNN-based attention blocks. Inspired by Chen et al. [4], which demonstrated a fully CNN-based model for image enhancement achieving state-of-the-art performance, we incorporated channel attention within the module and adapted it for time-series images. We integrated this modified CNN module into our structure. Our findings show that applying the CNN-based attention module after the transformer enhances PSNR and SSIM by better capturing local dependencies, resulting in sharper output frames compared to the original TMT. We believe that this integration effectively leverages the strengths of both CNNs and transformers, contributing to the simplicity and effectiveness of our network.

We have also conducted a qualitative analysis on how the number of input frames affects the turbulence mitigation task. Fig. 4 illustrates the model outputs with varying numbers of input frames, ranging from 1 to 20. Fig. 4 shows that increasing the number of input frames clearly enhances turbulence mitigation because the warping effect is closely tied to temporal variation.

4.4 Assessing the Generalization Performance

For the generalization assessment, we evaluate the models on the BVI-CLEAR [3], the real-world dataset [2,11], and a real-world text dataset [20]. The CLEAR dataset comprises static disturbed sequential frames with their corresponding ground truth images. This dataset mostly contains images with textual information. The objective of this experiment is to assess the models' capacity in exploiting textual information from turbulence-distorted videos. Fig. 3 illustrates



(a) Input

(b) TMT

(c) HATM (ours)

Fig. 3. Qualitative comparison on BVI-CLEAR [3] dataset. (a) represents the input while (b) shows the output of TMT [37]. (c) represents the output of the proposed HATM.

Table 2. Average inference PSNR and SSIM scores across dynamic video scenes for SOTA models.

Methods	PSNR	SSIM
TMT [37]	31.58	0.9209
RVRT [17]	32.38	0.9316
HATM (ours)	32.37	0.9332

Table 3. Quantitative comparison analysis of PSNR and SSIM scores obtained by different attention blocks on static dataset.

Methods	PSNR	SSIM
Vanilla CNN-based Attention Block	22.63	0.5821
Transformer+ Gated Feed-forward Block (TMT) [37]	23.64	0.7412
Our Transformer-CNN-based Attention Block (HATM)	24.03	0.7798

the results of TMT and our proposed model on the BVI-CLEAR dataset. The first column displays the first input frames of four different videos. The second column shows the corresponding output frames of TMT model. The last column shows the corresponding results of the proposed HATM model. The results demonstrate that HATM excels in reconstructing text content of turbulent video sequences. More specifically, for instance, the second row of Fig. 3 shows that HATM reconstructs the text "MAGAZINE" more effectively compared to TMT, highlighting its superior performance in preserving textual details amidst turbulence distortion.

We also evaluated the performance of the proposed model alongside TMT, using the real-world dataset [3] which includes static and dynamic real world turbulence-affected videos. The first row in Fig. 5 shows the models' outputs for the '*hill house*', a static scene, while the second row presents results for the '*moving car*', a dynamic video sequence. Our proposed model produces clearer and sharper images, outperforming TMT.

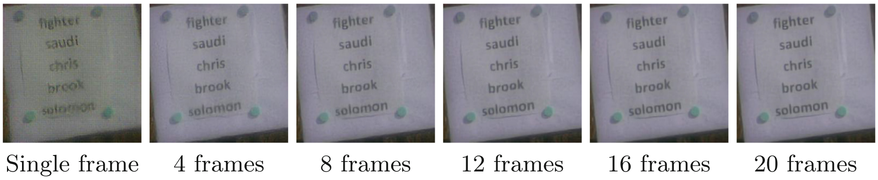
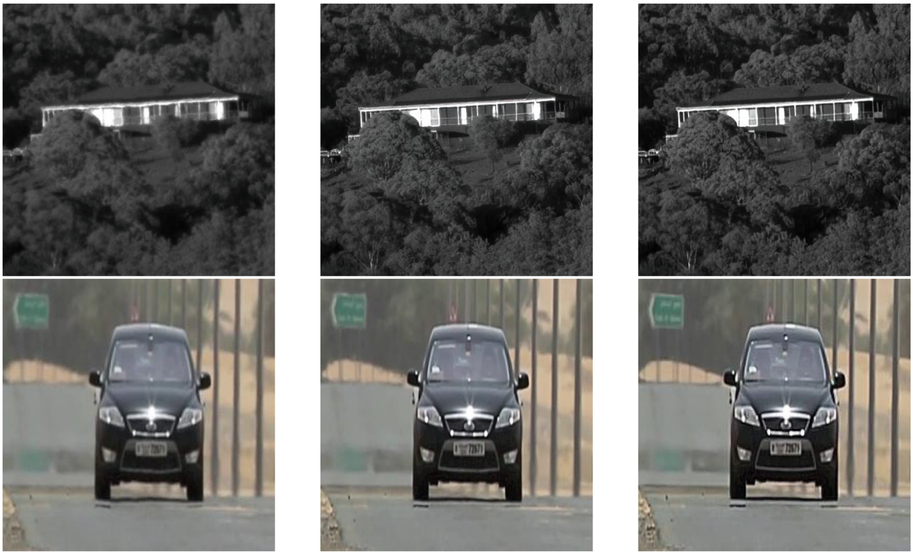
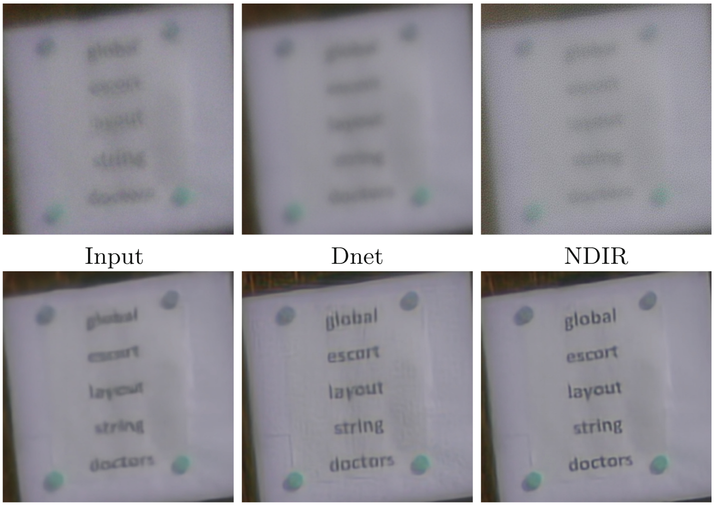


Fig. 4. Qualitative analysis of the proposed model's outputs with varying numbers of input frames. The outputs are displayed from left to right, corresponding to 1 to 20 input frames.



(a) Input (b) TMT (c) HATM (ours)

Fig. 5. Qualitative comparison on real world dynamic dataset. The first and second row demonstrate the performance of the models on a static '*hill house*' and dynamic '*moving car*' videos, respectively.



Input Dnet NDIR

VRT TMT HATM (ours)

Fig. 6. Qualitative analysis on real world text dataset. Dnet [14], NDIR [15], and TMT [37] are specifically designed for turbulence mitigation, whereas VRT [16] was originally developed for video restoration.

For further supporting our experiments, Fig. 6 illustrates qualitative comparisons on a real-world text dataset [20], involving recent turbulence mitigation and video restoration models, including Dnet [14], NDIR [15], VRT [16], and TMT [37]. All models were fine-tuned on our synthesized dataset, starting from their provided pre-trained weights.

5 Conclusion

In this study, we proposed a hybrid transformer-CNN-based attention model for video turbulence mitigation (HATM). HATM incorporated attention blocks, where each block integrated a combination of CNN-based and Transformer-based attention modules to effectively eliminate turbulence distortions in videos. To assess the performance of HATM, we conducted experiments on our simulated dataset as well as real data. The results of our study demonstrate the superior performance of HATM on our simulated video dataset over some recent SOTA models. We also evaluated the proposed model on the BVI-CLEAR, CLEAR, and real world turbulence-affected text datasets including static and dynamic videos to assess its generalization performance. The results demonstrate that our proposed model provides superior visual performance in handling turbulence mitigation in various scenes and delivers more clear and sharp frames. As a future work, we will train the model on larger datasets and enhance its capabilities to further improve the clarity of text images in the turbulence situation.

References

1. Agustsson, E., Timofte, R.: Ntire 2017 challenge on single image super-resolution: Dataset and study. In: Proceedings of the IEEE conference on computer vision and pattern recognition workshops. pp. 126–135 (2017)
2. Anantrasirichai, N.: Atmospheric turbulence removal with complex-valued convolutional neural network. Pattern Recogn. Lett. **171**, 69–75 (2023)
3. Anantrasirichai, N., Achim, A., Kingsbury, N.G., Bull, D.R.: Atmospheric turbulence mitigation using complex wavelet-based fusion. IEEE Trans. Image Process. **22**(6), 2398–2408 (2013)
4. Chen, L., Chu, X., Zhang, X., Sun, J.: Simple baselines for image restoration. In: European Conference on Computer Vision. pp. 17–33. Springer (2022)
5. Chen, Z., Zhang, Y., Gu, J., Kong, L., Yang, X., Yu, F.: Dual aggregation transformer for image super-resolution. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 12312–12321 (2023)
6. Chen, Z., Liang, Y., Du, M.: Attention-based broad self-guided network for low-light image enhancement. In: 2022 26th International Conference on Pattern Recognition (ICPR). pp. 31–38. IEEE (2022)
7. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
8. Fishbain, B., Yaroslavsky, L.P., Ideses, I.A.: Real time turbulent video perfecting by image stabilization and super-resolution. arXiv preprint arXiv:0704.3447 (2007)

9. Guo, H., Lu, T., Wu, Y.: Dynamic low-light image enhancement for object detection via end-to-end training. In: 2020 25th International Conference on Pattern Recognition (ICPR). pp. 5611–5618. IEEE (2021)
10. Han, K., Wang, Y., Chen, H., Chen, X., Guo, J., Liu, Z., Tang, Y., Xiao, A., Xu, C., Xu, Y., et al.: A survey on vision transformer. IEEE Trans. Pattern Anal. Mach. Intell. **45**(1), 87–110 (2022)
11. Jaiswal, A., Zhang, X., Chan, S.H., Wang, Z.: Physics-driven turbulence image restoration with stochastic refinement. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 12170–12181 (2023)
12. Karpathy, A., Toderici, G., Shetty, S., Leung, T., Sukthankar, R., Fei-Fei, L.: Large-scale video classification with convolutional neural networks. In: Proceedings of the IEEE conference on Computer Vision and Pattern Recognition. pp. 1725–1732 (2014)
13. Khan, S., Naseer, M., Hayat, M., Zamir, S.W., Khan, F.S., Shah, M.: Transformers in vision: A survey. ACM computing surveys (CSUR) **54**(10s), 1–41 (2022)
14. Khowaja, S.A., Lee, I.H., Yoon, J.: 2nd place solutions for ug2+ challenge 2022–dnet for mitigating atmospheric turbulence from images. arXiv preprint arXiv:2208.12332 (2022)
15. Li, N., Thapa, S., Whyte, C., Reed, A.W., Jayasuriya, S., Ye, J.: Unsupervised non-rigid image distortion removal via grid deformation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 2522–2532 (2021)
16. Liang, J., Cao, J., Fan, Y., Zhang, K., Ranjan, R., Li, Y., Timofte, R., Van Gool, L.: Vrt: A video restoration transformer. IEEE Transactions on Image Processing pp. 2171–2182 (2024)
17. Liang, J., Fan, Y., Xiang, X., Ranjan, R., Ilg, E., Green, S., Cao, J., Zhang, K., Timofte, R., Gool, L.V.: Recurrent video restoration transformer with guided deformable attention. Adv. Neural. Inf. Process. Syst. **35**, 378–393 (2022)
18. Liu, W., Peng, J., Yuan, H., Zhang, L., Cai, Z.: Mhrnet: A multi-stage image deblurring approach with high-resolution representation learning. In: 2023 International Joint Conference on Neural Networks (IJCNN). pp. 1–8. IEEE (2023)
19. Mao, Z., Chimitt, N., Chan, S.H.: Accelerating atmospheric turbulence simulation via learned phase-to-space transform. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 14759–14768 (2021)
20. Mao, Z., Jaiswal, A., Wang, Z., Chan, S.H.: Single frame atmospheric turbulence mitigation: A benchmark study and a new physics-inspired transformer model. In: European Conference on Computer Vision. pp. 430–446. Springer (2022)
21. Mohla, S., Pande, S., Banerjee, B., Chaudhuri, S.: Fusatnet: Dual attention based spectrospatial multimodal fusion network for hyperspectral and lidar classification. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. pp. 92–93 (2020)
22. Moran, S., McDonagh, S., Slabaugh, G.: Curl: Neural curve layers for global image enhancement. In: 2020 25th International Conference on Pattern Recognition (ICPR). pp. 9796–9803. IEEE (2021)
23. Park, J., Nah, S., Lee, K.M.: Recurrence-in-recurrence networks for video deblurring. arXiv preprint arXiv:2203.06418 (2022)
24. Parmar, N., Vaswani, A., Uszkoreit, J., Kaiser, L., Shazeer, N., Ku, A., Tran, D.: Image transformer. In: International conference on machine learning. pp. 4055–4064. PMLR (2018)
25. Rambhatla, S.S., Misra, I., Chellappa, R., Shrivastava, A.: Most: Multiple object localization with self-supervised transformers for object discovery. In: Proceedings

of the IEEE/CVF International Conference on Computer Vision. pp. 15823–15834 (2023)

26. Ronneberger, O., Fischer, P., Brox, T.: U-Net: Convolutional Networks for Biomedical Image Segmentation. In: Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F. (eds.) MICCAI 2015. LNCS, vol. 9351, pp. 234–241. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-24574-4_28
27. Selva, J., Johansen, A.S., Escalera, S., Nasrollahi, K., Moeslund, T.B., Clapés, A.: Video transformers: A survey. IEEE Transactions on Pattern Analysis and Machine Intelligence (2023)
28. Shivakumara, P., Jain, T., Pal, U., Surana, N., Antonacopoulos, A., Lu, T.: Text line segmentation from struck-out handwritten document images. Expert Syst. Appl. **210**, 118266 (2022)
29. Son, H., Lee, J., Lee, J., Cho, S., Lee, S.: Recurrent video deblurring with blur-invariant motion estimation and pixel volumes. ACM Transactions on Graphics (TOG) **40**(5), 1–18 (2021)
30. Souibgui, M.A., Biswas, S., Jemni, S.K., Kessentini, Y., Fornés, A., Lladós, J., Pal, U.: Docentr: An end-to-end document image enhancement transformer. In: 2022 26th International Conference on Pattern Recognition (ICPR). pp. 1699–1705. IEEE (2022)
31. Vint, D., Di Caterina, G., Soraghan, J., Lamb, R., Humphreys, D.: Analysis of deep learning architectures for turbulence mitigation in long-range imagery. In: Artificial Intelligence and Machine Learning in Defense Applications II. vol. 11543, p. 1154303. SPIE (2020)
32. Walha, A., Wali, A., Alimi, A.M.: Video stabilization for aerial video surveillance. Aasri Procedia **4**, 72–77 (2013)
33. Wang, L., Zhao, H., Guo, S., Mai, Y., Liu, S.: The adaptive compensation algorithm for small uav image stabilization. In: 2012 IEEE International Geoscience and Remote Sensing Symposium. pp. 4391–4394. IEEE (2012)
34. Wang, X., Yu, K., Wu, S., Gu, J., Liu, Y., Dong, C., Qiao, Y., Change Loy, C.: Esrgan: Enhanced super-resolution generative adversarial networks. In: Proceedings of the European conference on computer vision (ECCV) workshops. pp. 0–0 (2018)
35. Zamir, S.W., Arora, A., Khan, S., Hayat, M., Khan, F.S., Yang, M.H.: Restormer: Efficient transformer for high-resolution image restoration. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 5728–5739 (2022)
36. Zhang, J.B., Zhao, M.B., Yin, F., Liu, C.L.: Sequential transformer for end-to-end video text detection. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 6520–6530 (2024)
37. Zhang, X., Mao, Z., Chimitt, N., Chan, S.H.: Imaging through the atmosphere using turbulence mitigation transformer. IEEE Transactions on Computational Imaging **10**, 115–128 (2024)
38. Zhong, Z., Gao, Y., Zheng, Y., Zheng, B., Sato, I.: Real-world video deblurring: A benchmark dataset and an efficient recurrent neural network. Int. J. Comput. Vision **131**(1), 284–301 (2023)

DereflectFormer: Vision Transformers for Single Image Reflection Removal

Ao Wei[1,2], Hanbin Zhang[2], and Erhu Zhao[2,3](✉)

[1] ShanghaiTech University, Shanghai, China
weiao2022@shanghaitech.edu.cn

[2] Shanghai Innovation Center for Processor Technologies, Shanghai, China
zhanghanbin@shic.ac.cn

[3] Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China
zhaoerhu@ict.ac.cn

Abstract. In this paper, we address the challenge of single image reflection removal (SIRR), a crucial task in computer vision that involves eliminating undesirable reflections from images captured through glass surfaces. Current state-of-the-art methods typically rely on convolutional neural networks (CNNs) and often make certain assumptions about the appearance of reflections, which may not hold true in real-world scenarios. To overcome these limitations, we propose a novel Transformer-based approach, DereflectFormer, inspired by the Swin Transformer. Our architecture introduces a new module, the Depthwise Multi-Activation Feed-Forward Network (DMFN), which leverages depthwise convolution and a dual-stream ReLU-GELU activation function to enhance detail extraction capability. We also employ a synthetic dataset and a synthesis method for training, which allows our model to fully exploit the capabilities of Transformer architectures. Based on experimental results, we demonstrate that our approach performs better than state-of-the-art methods, providing more accurate and robust results in various real-world scenarios. Furthermore, our ablation studies reveal that each component of our architecture contributes significantly to its performance, offering valuable insights for future research in the field of single image reflection removal. The code and dataset is available at https://github.com/Agent76ow/DereflectFormer.

Keywords: Single Image Reflection Removal · Transformer-based Approach · Depthwise Multi-Activation Feed-Forward Network

1 Introduction

1.1 Background

In daily life, images captured through glass surfaces are frequently subject to undesirable reflections, which substantially degrade their quality and hinder the performance of various computer vision tasks, such as object recognition, semantic segmentation, and scene understanding. These challenges make single image

A. Antonacopoulos et al. (Eds.): ICPR 2024, LNCS 15321, pp. 257–274, 2025.
https://doi.org/10.1007/978-3-031-78305-0_17

reflection removal (SIRR) an essential low-level vision problem that has garnered significant attention from researchers in the computational photography and computer vision communities.

1.2 Challenges of Existing SIRR Methods

Current state-of-the-art SIRR methods [9,14,17] typically rely on convolutional neural networks (CNNs) and involve assumptions about the appearance of reflections, such as blurriness, ghosting cues [17,29] from thick glass, and differences in focus between reflection and transmission. However, these assumptions may not always be valid in real-world scenarios, where reflection appearances can be diverse and complex. Consequently, existing methods encounter difficulties in effectively handling and removing reflections with varying appearances from single input images.

Recently, there has been growing interest in leveraging the self-attention mechanism provided by Transformer [32] architectures for various computer vision tasks. While CNN-based approaches have rapidly advanced the state-of-the-art in low-level vision tasks, Transformer models, with their unique capabilities, present new possibilities for improved performance. Notably, Transformers have demonstrated their potential in high-level tasks through the success of Vision Transformers (ViT [8]), which outperforms the majority of CNN architectures. Following this success, many modified architectures [5,25,27,40] have been proposed to further adapt and refine the original model.

Despite these modifications primarily focusing on high-level tasks, several variations have recently emerged to address low-level challenges [4,15,24,36,41]. The inherent advantages of Transformer models, such as the ability to effectively capture long-range dependencies and contextual information within images, make them a promising choice for low-level tasks, such as reflection removal. As a result, this presents an opportunity to investigate how the strengths of Transformer models can be harnessed for reflection removal, potentially offering more accurate and robust results compared to traditional CNN-based approaches.

1.3 Proposed Solution: DereflectFormer

In this paper, we propose a novel Transformer-based approach dubbed DereflectFormer for SIRR, which is inspired by Swin Transformer [25]. Through our research, we have observed that the GELU activation function commonly employed in Vision Transformer models tends to underperform in reflection removal tasks. Additionally, the Multi-Layer Perceptron (MLP) layers struggle to effectively restore fine image details. To address these issues, we introduce a novel MLP module called Depthwise Multi-Activation Feed-Forward Network (DMFN). DMFN leverages depthwise convolution for extracting more hierarchical detail information and utilizes a dual-stream ReLU-GELU activation function to enhance detail extraction capability. The combined use of GELU [13] and ReLU [12] activation functions captures distinct types of nonlinear features, emphasizing their complementary effects. The depthwise separable convolution design of DMFN targets individual channels separately, reducing computational complexity while delivering substantial improvements in image restoration quality with a modest increase in the number of parameters.

Due to the need for a large amount of training data to fully exploit Transformer capabilities, the existing datasets are insufficient to unleash their full potential. Therefore, we employ a synthetic dataset and introduce a synthesis method (to be detailed in Sec. 4.2) to better support the training of our DereflectFormer model.

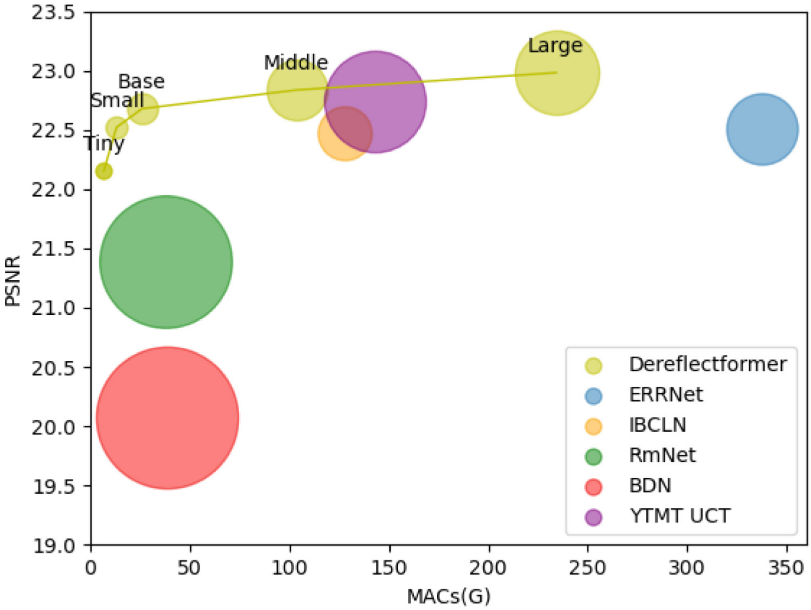


Fig. 1. Compared to other methods of image reflection removal, the DereflecFormer's performance is illustrated, with dot size representing the number of parameters and MAC values plotted on the **X**-axis.

1.4 Comparison and Contributions

In conclusion, the necessity for reflection removal in real-world images emphasizes the importance of continued research in single image reflection removal. Our work focuses on developing an approach that uses Transformer architectures named DereflectFormer, accompanied by the novel MLP module, DMFN. To further facilitate the learning of our DereflectFormer model, we make use of a synthesized dataset specifically tailored for this task. To demonstrate the efficiency of our solution, we have compared DereflectFormer with existing models. Fig. 1 demonstrates the superior performance of DereflectFormer in comparison to other image reflection removal models in terms of computational efficiency and performance metrics, underscoring the advantage of our method.

Building upon the previously discussed insights, we set forth our main contributions in this work as follows:

- We propose DereflectFormer, Transformer-based architecture for single-image reflection removal (SIRR). Using Swin Transformer blocks and a U-Net-like structure, our model offers effective image reflection removal across various real-world scenarios.
- We introduce the Depthwise Multi-Activation Feed-Forward Network (DMFN) module. Supplanting traditional GELU-based activations and MLP layers, DMFN enhances fine detail capture and restoration, improving overall reflection removal quality.
- To boost DereflectFormer's capabilities, we use a synthetic dataset and a synthesis method for training. This strategy provides diverse training samples, enhancing our model's applicability and generalization in real-world reflection removal scenarios.

2 Related Work

2.1 Image Reflection Removal

Early single image reflection removal methods were primarily non-learning based approaches. The majority of these methods relied on the defocused reflection assumption, which posits that reflections are usually more blurry compared to the true transmission [1,18,22,31,33]. By exploiting the blurry nature of defocused reflections, these methods attempt to suppress the reflection by leveraging image gradients. Another cue often utilized in this line of research is the ghosting cue, whereby multiple reflections are visible on the glass [17,29,31]. However, this cue is largely dependent on glass thickness, resulting in potential failure on thin glass surfaces.

With the development of deep learning techniques, learning-based methods have emerged as a dominant paradigm for single image reflection removal [9,42]. Generative Adversarial Networks (GANs) were employed by Wei *et al.* [38] and Ma *et al.* [26] to synthesize realistic reflections under guidance from real-world data. Another notable contribution was made by Kim *et al.* [16], who proposed a physics-based method to render reflections and mixed images, significantly improving the quality of training data. Despite these advancements, existing methods still face difficulties in perfectly removing reflections for a wide range of real-world data [2,9,30].

Various network structures have also been explored for reflection removal. Fan *et al.* [9] proposed CEILNet, a two-stage architecture that predicts edge maps before estimating the transmission layer. This two-stage approach was further developed by DMGN [10], Dong *et al.* [7], and RAGNet [23], who added the estimation of the reflection layer at the first stage. VGG-19 features were also combined with SIRR models by Zhang *et al.* [42], leading to a substantial improvement in results. Other advancements include recurrent neural network-based techniques for iterative reflection and transmission layer prediction [7,10,19].

2.2 Vision Transformers.

The success of CNNs in a variety of computer vision tasks has been well-documented. However, CNNs have certain limitations, such as restricted receptive fields and a large number of parameters, which can hinder their performance in more complex tasks. To address these issues, researchers have turned to the transformer architecture, which has shown promising results in the natural language processing domain with models such as BERT [6]. The self-attention mechanism in transformers facilitates the capturing of long-range dependencies and enhances interpretability. Vision Transformers (ViTs) have proven effective in computer vision tasks [8], demonstrating their potential as a fundamental building block in lieu of convolutions. The DereflectFormer is similar to Swin Transformer[25] and U-Net[28], but it offers several critical modifications to remove image reflections.

ViTs have found success not only in high-level visual tasks but also in low-level image restoration tasks, such as super-resolution and denoising. For instance, based on the Swin Transformer, SwinIR[24] is a strong baseline model for image restoration, has shown improved performance in these low-level tasks. Notably, Zamir *et al.* proposed an efficient Transformer model called Restormer [41], which excels in high-resolution image restoration tasks. Restormer achieves this by making key design choices in basic components, such as multi-head attention and feed-forward networks, allowing it to capture long-range pixel interactions and remain applicable to large images. Uformer [36], a novel model that capitalizes on Swin Transformer blocks to establish a U-Net-like network and incorporates depth-wise convolution (DWConv [11]) into the feed-forward network (FFN) in a semblance with the LocalViT strategy [21]. The Uformer ingeniously combines the ability of the Transformer to handle long-range dependencies with the strengths of U-Net in image processing tasks.

3 Method

3.1 Overall Pipeline

The DereflectFormer architecture, our proposed model, is specifically designed to efficiently tackle single-image reflection removal tasks. This architecture is built upon the popular Swin Transformer [25], with significant improvements to adapt to the unique challenges posed by reflection removal tasks. The loss function adopted by the network is the L_1 loss, which is commonly used for regression problems and works well for image reflection removal tasks due to its robustness against outliers.

The overall pipeline of our architecture is presented in Fig. 2. The pipeline begins with a degraded input image $\mathbf{I} \in \mathbb{R}^{\mathbf{H}\times\mathbf{W}\times\mathbf{3}}$, where $H \times W$ denotes the spatial dimension and 3 represents the number of channels in the RGB color space. The DereflectFormer initially applies a convolution operation to this input image to obtain low-level feature embeddings $\mathbf{F_0} \in \mathbb{R}^{\mathbf{H}\times\mathbf{W}\times\mathbf{C}}$, where C signifies the number of output channels. These shallow features F_0 are then passed through a symmetric encoder-decoder structure, where they are transformed into deep features $\mathbf{F_d} \in \mathbb{R}^{\mathbf{H\times W\times 2C}}$. There are multiple DereflectFormer blocks at each level of this encoder-decoder structure, with the number of blocks gradually increasing from the top to bottom levels to maintain computational efficiency.

The encoder part of the structure operates by progressively reducing the spatial size of the input while expanding the channel capacity, utilizing downsampling techniques. The decoder, on the other hand, starts with low-resolution latent features $\mathbf{F_l} \in \mathbb{R}^{\frac{\mathbf{H}}{\mathbf{4}}\times\frac{\mathbf{W}}{\mathbf{4}}\times\mathbf{4C}}$ and progressively recovers the high-resolution representations using upsampling techniques. To assist in the recovery process, the features from the encoder are concatenated through SK fusion with the features from the decoder via skip connections, a technique inspired by the U-Net architecture [28]. This process helps preserve the high-frequency details in the input image, which are crucial for reflection removal tasks.

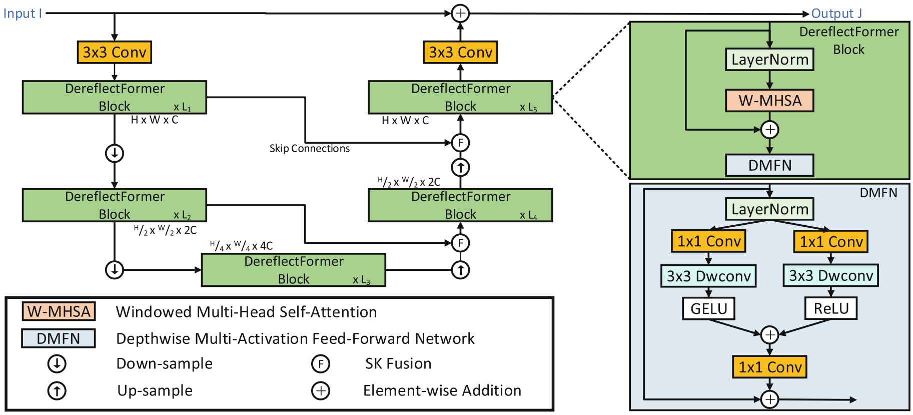


Fig. 2. The DereflectFormer is a five-stage U-Net-like architecture that incorporates several novel components specifically designed for single image reflection removal tasks. The architecture introduces a Depthwise Multi-Activation Feed-Forward Network (DMFN) module and SpatialLayerNorm into the DereflectFormer blocks to allow for controlled feature transformation and to better manage spatial structures in the image data. It also includes window-based multi-head self-attention (W-MHSA), enhancing the model's ability to aggregate spatial information. The SK fusion layer, an additional feature of this architecture, replaces traditional concatenate to improve feature fusion.

Finally, a convolution layer is employed on the refined features to generate a residual image $\mathbf{R} \in \mathbb{R}^{\mathbf{H\times W\times 3}}$. The degraded input image is then added to this residual image to obtain the restored image : $\hat{I} = I + R$. In addition to the above, our DereflectFormer architecture includes another components: the SK fusion layer, which replace the original concatenation fusion. The SK fusion layer is inspired by SKNet [20], and it is designed to adaptively recalibrate channel-wise feature responses for better feature fusion. Next, we present the details of the DMFN module.

3.2 Depthwise Multi-Activation Feed-Forward Network (DMFN) Module

In the Depthwise Multi-Activation Feed-Forward Network (DMFN) module, we introduce two fundamental modifications to the traditional feed-forward network (FN) [32] to improve representation learning: (1) multi-activation mechanism, and (2) depthwise convolutions. The architecture of our DMFN is depicted in Fig. 2, as seen on the right-hand side of the figure.

Feed-forward networks (FN) operate on each pixel location individually using two 1×1 convolutions to expand and reduce the feature channels (usually by a factor $\gamma = 4$) with a non-linearity applied in the hidden layer. In contrast, our DMFN module incorporates a multi-activation mechanism that combines the strengths of two different activation functions, ReLU [12] and GELU [13], in parallel paths. This design allows the module to capture a richer set of feature representations, as different activation functions may excel in modeling different aspects of the data.Depthwise convolutions [11] are also integrated into the DMFN module to encode information from spatially neighboring pixel positions, which is essential for learning local image structure and achieving effective restoration. Depthwise convolutions apply separate filters to each input channel, enabling the efficient learning of local features while keeping the computational burden low.

Given an input tensor $\mathbf{X} \in \mathbb{R}^{\mathbf{H}\times\mathbf{W}\times\mathbf{C}}$, the DMFN module can be formulated as follows:

$$\hat{\mathbf{X}} = \mathbf{MLA}(\mathrm{LN}(\mathbf{X})) + \mathbf{X}. \tag{1}$$

In Eqn. 1, the DMFN module consists of a MultiActivation (MLA) function, a residual connection, and layer normalization (LN) [3]. The MLA function, which is a key component in the DMFN module, combines the advantages of two different activation functions, ReLU and GELU. This mechanism enhances representation learning for reflection removal tasks by allowing the module to capture a richer set of feature representations.

The detailed operation of the MLA function unfolds as follows:

$$\mathbf{MLA}(\mathbf{X}) = \phi(\mathrm{ReLU}(\epsilon(\mathbf{X})) + \mathrm{GELU}(\epsilon(\mathbf{X}))). \tag{2}$$

In the above equation, ϕ denotes a 1×1 convolution operation that reduces the dimensionality of the output, and ϵ represents a sequence of operations: a 1×1 convolution that expands the dimensionality, followed by a 3×3 depthwise convolution for feature extraction. This function first expands the dimensionality of the input through a 1×1 convolution operation, represented by ϵ. It then applies a 3×3 depthwise convolution to extract spatial information from the input, before applying the ReLU and GELU activation functions in parallel. The outputs of these two activation functions are then summed, which allows the function to capture a richer set of feature representations. The summed output is finally reduced back to the original dimensionality through a 1×1 convolution operation, represented by ϕ.

The channel expansion factor in the DMFN module varies between different DereflectFormer blocks in the network. Specifically, the pattern expansion ratio, denoted as γ, is defined as [2, 4, 4, 2, 2], meaning that the first and last blocks double the number of channels, while the second and third quadruple them. This design allows the DMFN module to control the information flow through different hierarchical levels in our pipeline, and maintain a similar number of parameters and computational burden compared to traditional feed-forward networks (FN [32]). This way, the DMFN module is able to capture both fine-grained details and high-level contextual information effectively and efficiently.

3.3 SpatialLayerNorm and W-MHSA Components

We integrate SpatialLayernorm and W-MHSA components into our DereflectFormer architecture to further improve its performance. These components are designed to better address the unique challenges presented by single image reflection removal tasks. The SpatialLayerNorm is a variant of the traditional Layer Normalization (LayerNorm [3]). The main difference lies in the dimensions used for calculating the mean and standard deviation.

In the original LayerNorm, these statistics are computed along the last dimension of the input (dim = -1). This means that normalization is performed independently for each feature in the input, making it suitable for handling structured 2D inputs, such as the outputs of fully connected layers.

In contrast, SpatialLayerNorm computes these statistics along the channel, height, and width dimensions of the input (dim = (1, 2, 3)). This means that normalization is performed independently for each channel. This adjustment allows SpatialLayerNorm to better handle 4D inputs with spatial structure, which are typical in our application (i.e., image reflection removal tasks). The 4D inputs refer to a batch of color images, where the dimensions represent (batch size, channels, height, width). The spatial normalization operation enables the model to effectively account for the variations in different spatial locations across the channel dimension, enhancing the capability of the model to capture spatially varying patterns in the input image.

The W-MHSA (Window-based Multi-Head Self-Attention) component is a variant of the traditional MHSA that aims to enhance the capability of the model to aggregate spatial information. Inspired by the Swin Transformer [25], we first partition the input feature map X into several non-overlapping windows and then apply the W-MHSA within each window.

Given an input feature map $\mathbf{X} \in \mathbb{R}^{\mathbf{b}\times\mathbf{h}\times\mathbf{w}\times\mathbf{c}}$, where b, h, w, and c denote the batch size, height, width, and number of channels, respectively, we project $\mathbf{X}$ into Q, K, and V (query, key, and value) using linear layers. Let $Q, K, V \in \mathbb{R}^{b\times l\times d}$ correspond to a single window & header, where l is the number of tokens in a window and d is the dimension. The self-attention is computed by :

$$\text{Attention}(Q, K, V) = \text{Softmax}\left(\frac{QK^T}{\sqrt{d}} + B\right)V, \tag{3}$$

where B denotes the relative position bias, with a linear layer positioned to transform the attention's yield.

In addition to the attention mechanism, we also apply a convolution operation on V to aggregate information from the neighboring pixels without considering window partitioning. The methodology for aggregating spatial information is therefore designed as follows:

$$\begin{aligned}\text{Aggregation}(Q, K, V) = \text{Softmax}\left(\frac{QK^T}{\sqrt{d}} + B\right)V \\ + \text{Conv}(\hat{V}),\end{aligned} \tag{4}$$

where $\hat{V} \in \mathbb{R}^{b \times h \times w \times c}$ signifies V prior to window partitioning, and Conv($\cdot$) can be either a depthwise convolution denoted as DWConv, or a ConvBlock, which is a sequence of Conv-ReLU-Conv operations. This aggregation scheme allows us to capture both intra-window and inter-window dependencies, resulting in a more effective representation learning process for reflection removal tasks.

4 Experiment and Analysis

4.1 Experimental Setup

We experiment with five DereflectFormer variants (denoted as -T, -S, -B, -M and -L for tiny, small, basic, middle, and large, respectively). The attention ratio in this context refers to the proportion of blocks that incorporate multi-head self-attention (MHSA), and we place the blocks containing MHSA at the end of each stage. For the three smaller models (-T, -S, -B, -M), we use depthwise convolution (DWConv) with a kernel size of 5 as the parallel convolutions. Since DWConv is an operation with low computational cost but high memory access cost, we use a ConvBlock with a kernel size of 3 for the large model (-L).

During the training phase, images are randomly cropped into patches of 224 × 224. For training diverse variants, we establish separate mini-batch sizes, i.e., 32, 16, 16, 16, 8 for −T, −S, −B, −M, −L, respectively. Referring to the linear scaling rule, we set the initial learning rate to 2, 1, 1, 1, 0.5×10^{-4} for −T, −S, −B, −M, −L, respectively. We train the models utilizing the AdamW optimizer and a cosine annealing scheme, where the learning rate is gradually lowered from the initial value down to 2, 1, 1, 1, 0.5×10^{-6}. To evaluate the overhead, we utilize the number of parameters (#Param) and multiply-accumulate operations (MACs). MACs are measured on 224 × 224 images.

4.2 DataSet

To train our DereflectFormer model, we create a synthetic dataset and employ a synthesis method. These resources enable the model to learn complex reflection patterns in a controlled environment, ultimately improving its performance on real-world images. We use a total of 329 images from the CID [35] dataset, along with 90 real images and 13700 synthetic images from Zhang *et al*. [42] to train our DereflectFormer model.

The synthetic images were generated using a novel method, distinct from that proposed by Zhang et al. [42]. Although their dataset was employed for the image synthesis, but their synthetic data images are biased and don't work well for my model training, the method I developed is entirely new and differs from the approach used by Zhang et al. To synthesize an image, two images are selected: A background image **B** and a reflection image **R**. Then, both **B** and **R** images are converted to NumPy arrays and normalized to a range between 0 and 1. Following this, a Gaussian kernel is created with a magnitude proportional to the kernel size and sigma values (which are randomized from predefined ranges for each iteration). This kernel is then convolved with the **R** image to produce a blurred version of the reflection image dubbed $\mathbf{R}_{blur}$. Fig. 3 visually illustrates the transformation process, presenting the original background image **B**, reflection image **R**, the blurred reflection image $\mathbf{R}_{blur}$, and the final synthetic image **M**. The final synthetic image **M** is created by adding **B** and $\mathbf{R}_{blur}$. If the maximum value in **M** exceeds 1 (the maximum value in the normalized range), a correction is applied to $\mathbf{R}_{blur}$ and **M** to ensure all values are restricted within the normalized range. The created synthetic image **M** is then saved and can be utilized for further testing or training a model for image reflection removal.

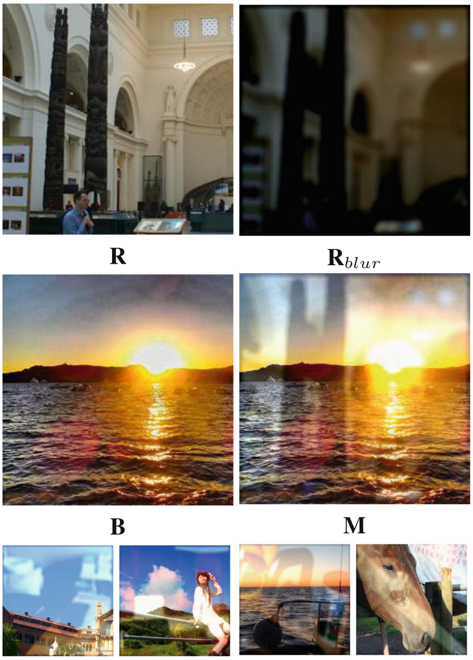


Fig. 3. This figure illustrates the process of synthesizing images. Here, **R** represents the reflection image that undergoes Gaussian blurring to yield $\mathbf{R}_{blur}$, a blurred version of the original reflection. **B** refers to the background image, and **M** signifies the final synthesized image, produced by combining **B** and $\mathbf{R}_{blur}$. Additionally, four examples of synthesized reflection images are presented at the bottom of the figure.

4.3 Results

In this subsection, we present the results of our experiments. We compare the performance of our proposed DereflectFormer model with state-of-the-art reflection removal methods on various datasets. We also provide quantitative analyses of the results to establish the effectiveness of our Dereflectformer. We quantitatively compare the performance of our DereflectFormer and some other reflection removal methods, including BDN [39], RmNet [38], IBCLN [19], ERRNet [37] and YTMT-UCT [14], on five real-world dataset, involving Real20 [42], three subsets of SIR^2 [34], and CID [35]. In Table 1, the average PSNR(Peak Signal-to-Noise Ratio), SSIM(Structural Similarity Index Measure), and overhead metrics for these models across five datasets are displayed. We provide a quantitative

Table 1. Single image reflection removal results. Quantitative comparison (average PSNR/SSIM on the 5 testing datasets and computational overhead) with different methods for single image reflection removal.

Method	AVG PSNR	AVG SSIM	Overhead	
			#Param	MACs
BDN	20.068	0.801	75.1626M	38.7042G
RmNet	21.386	0.83	65.4328M	37.4309G
IBCLN	22.478	0.843	10.8042M	127.3909G
ERRNet	22.508	0.832	18.9534M	337.2426G
YTMT-UCT	22.738	0.847	38.4485M	142.8246G
DereflectFormer-T	22.152	0.85	0.9434M	6.7481G
DereflectFormer-S	22.52	0.858	1.7978M	13.1525G
DereflectFormer-B	22.678	0.859	3.5435M	25.9384G
DereflectFormer-M	22.836	0.864	13.7527M	103.8233G
DereflectFormer-L	**22.982**	**0.865**	26.4934M	234.1255G

Table 2. Quantitative comparison on the testing datasets of different reflection removal methods. The best results are highlighted in red and the second best results in blue.

Method	Real20		CID		Wild		Solid		Postcard	
	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
BDN	18.18	0.717	17.55	0.738	21.35	0.846	22.66	0.852	20.60	0.851
RmNet	18.86	0.735	18.64	0.772	24.71	0.900	23.78	0.875	20.94	0.866
IBCLN	20.43	0.755	18.82	0.791	24.13	0.899	24.51	0.886	24.50	0.885
ERRNet	21.96	0.786	18.99	0.740	25.18	0.886	24.78	0.890	21.63	0.856
YTMT-UCT	21.21	0.773	19.51	0.787	25.48	0.909	24.84	0.895	22.65	0.872
DereflectFormer-T	20.20	0.772	21.39	0.816	24.06	0.908	23.85	0.896	21.26	0.858
DereflectFormer-S	20.70	0.789	21.85	0.825	24.44	0.914	24.43	0.903	21.18	0.861
DereflectFormer-B	21.07	0.792	22.26	0.828	25.03	0.918	24.28	0.903	20.75	0.853
DereflectFormer-M	21.27	0.798	22.37	0.836	24.40	0.918	24.38	0.906	21.76	0.860
DereflectFormer-L	21.31	0.804	22.77	0.840	25.00	0.919	24.85	0.912	20.98	0.852

comparison between the performance of DereflectFormer and several other models, with the results presented in Table 2. Results where DereflectFormer surpasses the other models are highlighted in bold for clear emphasis. The visual comparison is also conducted in Fig. 4.

In Table 2, the performance on the postcard dataset is notably worse compared to other datasets. This is because the postcard dataset is a synthetic dataset that includes extensive high-saturation reflections. Our model underperforms on this dataset due to the lack of training data with similar characteristics, which leads to more frequent failures in handling these challenging reflections.

4.4 Ablation Studies

In this section, we provide an in-depth ablation study to clarify the contributions of various components of our proposed DereflectFormer architecture in the task of single-image reflection removal. The detailed quantitative results are illustrated in Fig. 5 and Table 3.

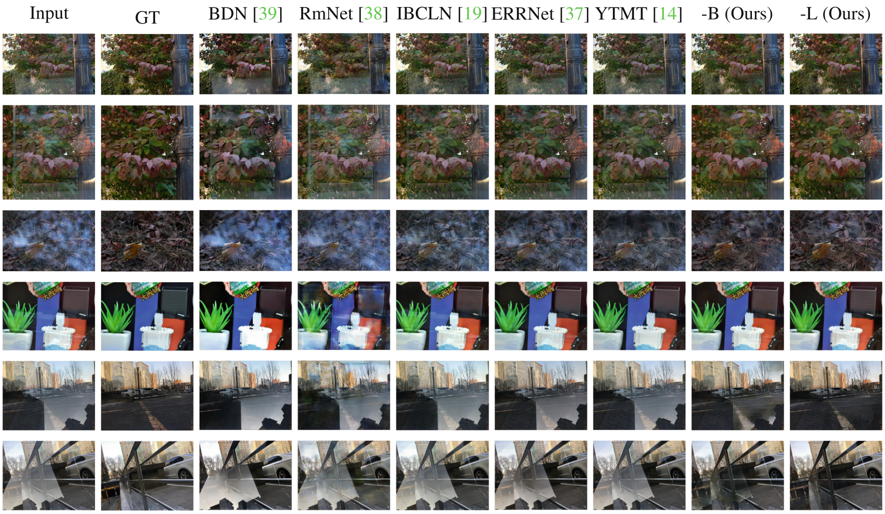


Fig. 4. Qualitative comparison of image reflection removal methods on real-world images. The images are obtained from 'Real20 [42]' (Rows 1-3), 'Solid [34]' (Row 4) and CID [35] (Rows 5-6). The first column is the reflect images and the last column is the corresponding ground truth.

To assess the impact of SpatialLayerNorm in our DereflectFormer, we replaced it with standard LayerNorm and observed a notable drop in performance. Standard LayerNorm, which normalizes features independently and is suitable for 2D data, is less effective for our 4D image data (batch size, channels, height, width). This change resulted in lower PSNR values, illustrating that SpatialLayerNorm's channel-wise normalization is crucial for handling spatial variations in reflection removal tasks. The experiment confirms the importance of SpatialLayerNorm in enhancing our model's ability to reduce reflections effectively.

We remove the Depthwise Convolution (DwConv) from the DereflectFormer-B model. The DwConv is designed to replace regular convolutions, while convolutions results in a slight improvement in performance, leads to a substantial increase in the number of model parameters and computational cost. Thus, our findings suggest that the Conv module may be less suitable for practical applications due to its demanding resource requirements.

We ablate the SKfusion module, which is devised to replace the traditional concatenate operation. Our results demonstrate that the omission of this component leads to a performance decrease, underlining its critical role in information fusion and feature intermixing.

We scrutinize the dual activation function approach implemented in the DMFN module, paying particular attention to the ReLU and GELU functions. Through trials with models employing either ReLU or GELU solely, the empirical evidence advocates that the dual usage of these activation functions significantly betters the model performance. This denotes that multi-activation function strategy substantially amplifies the model's representation capabilities and its adaptive performance across varied datasets.

Expanding on the activation function analysis, we introduce two supplementary ablations: "Double GELUs" and "Double ReLUs", both of which incorporate doubled instances of GELU and ReLU, respectively, within the DMFN module. Remarkably, models harnessing the double activation approach fail to outperform the original DMFN configuration, emphasizing that a mere increase

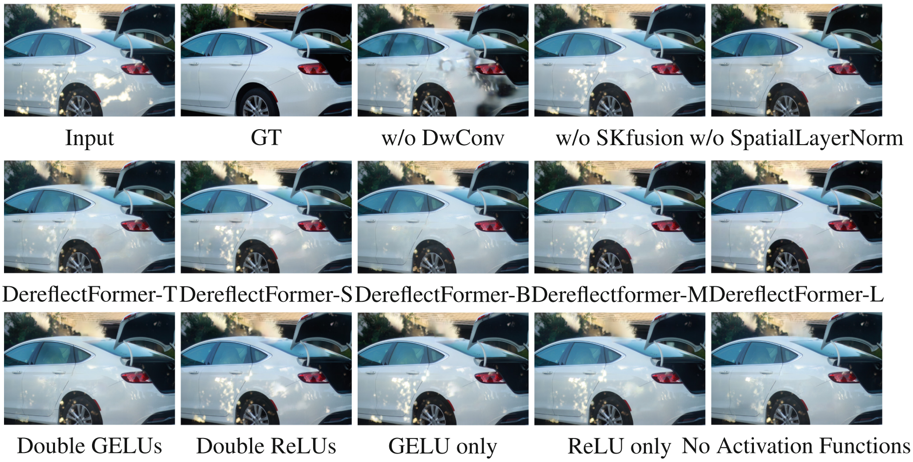


Fig. 5. Visual comparison of reflection removal results using DereflectFormer-B and its ablated variants. Each column showcases the effects of ablation on image quality and reflection suppression efficacy.

Table 3. Ablation experiments for Dereflectformer-B. PSNR is calculated as an average PSNR across 5 testing datasets, and #Param and MACs represent the computational overhead.

Method	PSNR	SSIM	#Param	MACs
DereflectFormer-B	22.678	0.859	3.5435M	25.9384G
w/o SpatialLayerNorm	22.188	0.858	3.5435M	25.9384G
w/o DwConv	22.784	0.859	7.1897M	55.7164G
w/o SKfusion	22.276	0.858	3.5423M	25.9365G
GELU only	21.970	0.854	2.5078M	18.1350G
ReLU only	22.310	0.847	2.5078M	18.1350G
Double GELUs	22.200	0.856	3.5435M	25.9384G
Double ReLUs	22.120	0.851	3.5435M	25.9384G
No Activation Functions	22.080	0.851	3.5435M	25.9384G

in activation function frequency does not correlate with enhanced performance. These instances reaffirm that strategic diversity, rather than redundancy, in activation functions is integral to achieving optimal model results.

Furthermore, the "No Activation Functions" scenario, whereby our model employs dual residual connections without any activation functions, serves to further manifest the indispensability of the activation role within the DMFN. Ablating the activation function results in a model with notably poorer capabilities in representation learning and non-linear transformation handling. This underscores the non-trivial role that activation functions play in the overall neural network architecture and their influence on model performance.

These comprehensive studies collectively amplify the understanding of each component's function, and inform the fine-tuning of our architecture for enhanced performance in reflection removal applications.

5 Discussion and Conclusion

5.1 Discussion

Through extensive experimentation and detailed analy- sis carried out in this work, we have gained a deeper understanding of the challenges associated with single image reflection removal, as well as potential solutions to tackle these problems. Our proposed DereflectFormer model and the introduction of the DMFN module have demonstrated promising results.

The data from our experiments show that DereflectFormer notably excels over other methods of the same period with a lower overhead. Fig. 1 shows the comparison of DereflectFormer with other image reflection removal methods on the five test dataset. Our small model defeats the IBCLN [19] with only 16.6% #Param and 10.3% computational cost. Our middle model outperforms in all fronts than the previous state-of-the-art method , YTMT-UCT [14]. Our large model performs better and still has a lower number of parameters than the previous state-of-the-art method, substantially outperforming contemporaneous methods.

While significant progress has been made, we acknowledge that there are still many unknowns and potential areas for improvement. For example, we plan to continue refining our model by exploring different activation functions, further optimizing parameters, and testing on more diverse datasets. The synthetic dataset and synthesis method we used for training also call for robust and comprehensive evaluation.

One critical limitation of our model is its dependency on large datasets for effective training. Insufficient or low-quality data can lead to overfitting or underfitting, thereby compromising the model's generalizability and robustness. Furthermore, the synthetically generated training data might not adequately represent the complexities and fine distinctions of real-world environments, thereby limiting the model's applicability to practical challenges. Future research should focus on curating high-quality datasets to mitigate these issues and enhance the model's reliability and performance.

5.2 Conclusion

This paper has presented a novel Transformer-based approach, DereflectFormer, specifically designed for the challenging task of single image reflection removal. The innovative DMFN module, coupled with the use of a synthetic dataset for its training, has boosted the performance of the DereflectFormer architecture.

Acknowledgements. This work is supported by the Shanghai Innovation Center for Processor Technologies.

References

1. Agrawal, A., Raskar, R., Nayar, S.K., Li, Y.: Removing photography artifacts using gradient projection and flash-exposure sampling. In: ACM SIGGRAPH 2005 Papers, pp. 828–835 (2005)
2. Arvanitopoulos, N., Achanta, R., Susstrunk, S.: Single image reflection suppression. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4498–4506 (2017)
3. Ba, J.L., Kiros, J.R., Hinton, G.E.: Layer normalization. arXiv preprint arXiv:1607.06450 (2016)
4. Chen, L., Chu, X., Zhang, X., Sun, J.: Simple baselines for image restoration. In: Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part VII. pp. 17–33. Springer (2022)
5. Choromanski, K., Likhosherstov, V., Dohan, D., Song, X., Gane, A., Sarlos, T., Hawkins, P., Davis, J., Mohiuddin, A., Kaiser, L., et al.: Rethinking attention with performers. arXiv preprint arXiv:2009.14794 (2020)
6. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018)

7. Dong, Z., Xu, K., Yang, Y., Bao, H., Xu, W., Lau, R.W.: Location-aware single image reflection removal. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 5017–5026 (2021)
8. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
9. Fan, Q., Yang, J., Hua, G., Chen, B., Wipf, D.: A generic deep architecture for single image reflection removal and image smoothing. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV) (Oct 2017)
10. Feng, X., Pei, W., Jia, Z., Chen, F., Zhang, D., Lu, G.: Deep-masking generative network: A unified framework for background restoration from superimposed images. IEEE Trans. Image Process. **30**, 4867–4882 (2021)
11. Fran, C., et al.: Deep learning with depth wise separable convolutions. In: IEEE conference on computer vision and pattern recognition (CVPR) (2017)
12. Glorot, X., Bordes, A., Bengio, Y.: Deep sparse rectifier neural networks. In: Proceedings of the fourteenth international conference on artificial intelligence and statistics. pp. 315–323. JMLR Workshop and Conference Proceedings (2011)
13. Hendrycks, D., Gimpel, K.: Gaussian error linear units (gelus). arXiv preprint arXiv:1606.08415 (2016)
14. Hu, Q., Guo, X.: Trash or treasure? an interactive dual-stream strategy for single image reflection separation. Adv. Neural. Inf. Process. Syst. **34**, 24683–24694 (2021)
15. Ji, H., Feng, X., Pei, W., Li, J., Lu, G.: U2-former: A nested u-shaped transformer for image restoration. arXiv preprint arXiv:2112.02279 (2021)
16. Kim, S., Huo, Y., Yoon, S.E.: Single image reflection removal with physically-based training images. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5164–5173 (2020)
17. Lei, C., Jiang, X., Chen, Q.: Robust reflection removal with flash-only cues in the wild. arXiv preprint arXiv:2211.02914 (2022)
18. Levin, A., Weiss, Y.: User assisted separation of reflections from a single image using a sparsity prior. IEEE Trans. Pattern Anal. Mach. Intell. **29**(9), 1647–1654 (2007)
19. Li, C., Yang, Y., He, K., Lin, S., Hopcroft, J.E.: Single image reflection removal through cascaded refinement. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3565–3574 (2020)
20. Li, X., Wang, W., Hu, X., Yang, J.: Selective kernel networks. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 510–519 (2019)
21. Li, Y., Zhang, K., Cao, J., Timofte, R., Van Gool, L.: Localvit: Bringing locality to vision transformers. arXiv preprint arXiv:2104.05707 (2021)
22. Li, Y., Brown, M.S.: Exploiting reflection change for automatic reflection removal. In: Proceedings of the IEEE international conference on computer vision. pp. 2432–2439 (2013)
23. Li, Y., Liu, M., Yi, Y., Li, Q., Ren, D., Zuo, W.: Two-stage single image reflection removal with reflection-aware guidance. Applied Intelligence pp. 1–16 (2023)
24. Liang, J., Cao, J., Sun, G., Zhang, K., Van Gool, L., Timofte, R.: Swinir: Image restoration using swin transformer. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 1833–1844 (2021)

25. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 10012–10022 (2021)
26. Ma, D., Wan, R., Shi, B., Kot, A.C., Duan, L.Y.: Learning to jointly generate and separate reflections. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 2444–2452 (2019)
27. Messina, N., Falchi, F., Esuli, A., Amato, G.: Transformer reasoning network for image-text matching and retrieval. In: 2020 25th International conference on pattern recognition (ICPR). pp. 5222–5229. IEEE (2021)
28. Ronneberger, O., Fischer, P., Brox, T.: U-Net: Convolutional Networks for Biomedical Image Segmentation. In: Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F. (eds.) MICCAI 2015. LNCS, vol. 9351, pp. 234–241. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-24574-4_28
29. Shih, Y., Krishnan, D., Durand, F., Freeman, W.T.: Reflection removal using ghosting cues. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3193–3201 (2015)
30. Souibgui, M.A., Biswas, S., Jemni, S.K., Kessentini, Y., Fornés, A., Lladós, J., Pal, U.: Docentr: An end-to-end document image enhancement transformer. In: 2022 26th International Conference on Pattern Recognition (ICPR). pp. 1699–1705. IEEE (2022)
31. Sun, C., Liu, S., Yang, T., Zeng, B., Wang, Z., Liu, G.: Automatic reflection removal using gradient intensity and motion cues. In: Proceedings of the 24th ACM international conference on Multimedia. pp. 466–470 (2016)
32. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in neural information processing systems **30** (2017)
33. Wan, R., Shi, B., Duan, L.Y., Tan, A.H., Gao, W., Kot, A.C.: Region-aware reflection removal with unified content and gradient priors. IEEE Trans. Image Process. **27**(6), 2927–2941 (2018)
34. Wan, R., Shi, B., Duan, L.Y., Tan, A.H., Kot, A.C.: Benchmarking single-image reflection removal algorithms. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 3922–3930 (2017)
35. Wang, C., Xu, D., Wan, R., He, B., Shi, B., Duan, L.Y.: Background scene recovery from an image looking through colored glass. IEEE Transactions on Multimedia (2022)
36. Wang, Z., Cun, X., Bao, J., Zhou, W., Liu, J., Li, H.: Uformer: A general u-shaped transformer for image restoration. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 17683–17693 (2022)
37. Wei, K., Yang, J., Fu, Y., Wipf, D., Huang, H.: Single image reflection removal exploiting misaligned training data and network enhancements. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8178–8187 (2019)
38. Wen, Q., Tan, Y., Qin, J., Liu, W., Han, G., He, S.: Single image reflection removal beyond linearity. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3771–3779 (2019)
39. Yang, J., Gong, D., Liu, L., Shi, Q.: Seeing deeply and bidirectionally: A deep learning approach for single image reflection removal. In: Proceedings of the european conference on computer vision (ECCV). pp. 654–669 (2018)

40. Yuan, L., Chen, Y., Wang, T., Yu, W., Shi, Y., Jiang, Z.H., Tay, F.E., Feng, J., Yan, S.: Tokens-to-token vit: Training vision transformers from scratch on imagenet. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 558–567 (2021)
41. Zamir, S.W., Arora, A., Khan, S., Hayat, M., Khan, F.S., Yang, M.H.: Restormer: Efficient transformer for high-resolution image restoration. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5728–5739 (2022)
42. Zhang, X., Ng, R., Chen, Q.: Single image reflection separation with perceptual losses. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4786–4794 (2018)

LK-Net: Efficient Large Kernel ConvNet for Document Enhancement

Qijun Shi[1], Hongjian Zhan[1](✉), Yangfu Li[1], Weijun Zou[2], Huasheng Li[2], Umapada Pal[3], and Yue Lu[1]

[1] Shanghai Key Laboratory of Multidimensional Information Processing, East China Normal University, Shanghai 200241, China
qijun.shi@stu.ecnu.edu.cn, ecnuhjzhan@foxmail.com, ylu@cs.ecnu.edu.cn

[2] Shanghai Hex Information Technology Co., Ltd., Shanghai, China
{wj.zou,colin.lee}@hexinfo.cn

[3] Indian Statistical Institute Kolkata, Kolkata, India
umapada@isical.ac.in

Abstract. Various types of degradation in document images, such as blurring, shadow, and physical wear and tear, significantly impact the effectiveness of downstream tasks in multimedia applications. The need for document image enhancement arises from the urgent need to improve the legibility and quality of these images, which are integral for accurate Optical Character Recognition(OCR), information retrieval, document analysis, etc. This paper introduces a novel and simple approach employing Large Kernel Convolutional Networks (ConvNets) for document image enhancement, capitalizing on their ability to encapsulate expansive contextual information to improve image quality. Extensive experimental evaluations across multiple benchmarks have demonstrated that our method achieves state-of-the-art (SOTA) while maintaining low computational cost. Code and pre-trained models are available at https://github.com/qijunshi/LKNet.

Keywords: Document Enhancement · Document Deblurring · Document Binarization

1 Introduction

In the realm of multimedia, document images play a pivotal role, serving as the foundation for a plethora of applications ranging from digital archiving and retrieval to automated content analysis. However, the performance of downstream tasks such as OCR[17], document classification[1], and content extraction[30] is significantly impeded by prevalent issues like document blur and noise. These challenges stem from a variety of factors, including poor scanning quality, physical wear and tear of documents, and environmental conditions affecting the document at the time of digitization.

In response to these challenges, a diverse array of document enhancement models[21,38] have emerged, leveraging the advancements in deep learning to

A. Antonacopoulos et al. (Eds.): ICPR 2024, LNCS 15321, pp. 275–290, 2025.
https://doi.org/10.1007/978-3-031-78305-0_18

enhance the quality and readability of degraded document images. Among these, ConvNets have been at the forefront, attributed to their proficiency in capturing spatial hierarchies of features within images. ConvNet-based models[10] provide substantial improvements in document enhancement. However, they often struggle with understanding the global context of document images due to their local receptive fields. To mitigate this issue, integration of multi-scale feature extraction methods is necessary, which can expand the receptive field and enhance the model's ability to comprehend broader contexts.

Generative Adversarial Networks (GANs)[8] introduce a novel paradigm where two neural networks, the generator and the discriminator, are pitted against each other in a game-theoretic framework. This approach[33] has proven effective for document enhancement, particularly in generating sharp, high-resolution images. GANs are adept at filling in missing information in document images, making them highly suitable for more severe cases of document deterioration. Nevertheless, GAN-based models can sometimes generate unrealistic artifacts or fail to preserve the fidelity of original content, posing challenges for their application in sensitive domains.

Transformer[35], originally designed for natural language processing, have recently been adapted for image processing tasks, including image classification[40], object detection[15], and document enhancement[32]. Their ability to capture long-range dependencies within the data makes them particularly suitable for understanding the global context of document images. Transformer-based models have shown promising results in capturing the intricate details of text and layout in document images. However, their high computational demand and the necessity for large-scale datasets for training are notable drawbacks.

ConvNets have predominantly utilized small kernels, such as 3×3 or 5×5, motivated by the belief that small, local features aggregate to form more complex patterns through the depth of the network. Small kernels often struggle to capture the broader context and the structural integrity of text and graphics in document images, leading to suboptimal restoration outcomes where the global coherence of the document layout is essential. Large kernels, extending beyond the conventional sizes, enable the network to encompass a large receptive field of view in a single convolutional step. This capacity allows for a more comprehensive understanding of the document image's global structure, facilitating the restoration of text and imagery with greater accuracy and fidelity. Moreover, large kernel ConvNets can achieve superior performance with fewer layers and parameters compared to their small-kernel counterparts, thanks to their efficiency in capturing spatial relationships over larger areas. This efficiency not only improves the quality of the restored images but also reduces the computational complexity and training time, making large kernel ConvNets a more practical solution for document enhancement.

The demand for generating high-resolution images has made the application of attention mechanisms in the field of image restoration challenging due to their high computational cost. The success of Transformers lies in their ability to easily capture global features and model the relationships between global pixels. Research[5] indicates that the receptive fields of networks with deep, small convolutional kernels are not as large as commonly believed. Conversely, networks with shallow, large convolutional kernels can have significantly large receptive fields, allowing them to capture global features similar to Transformers while requiring less computational cost. This realization points towards an innovative direction in document enhancement, where the adoption of large convolutional kernels offers a promising alternative to attention mechanisms, providing a balance between computational efficiency and the ability to capture and utilize global image features effectively.

Our contributions can be summarized in three points:

- We have successfully introduced large convolutional kernels into the field of document enhancement, achieving SOTA performance while maintaining low computational costs. This accomplishment demonstrates that ConvNets still hold the potential to rival Transformers in the document enhancement domain, highlighting the efficiency and effectiveness of large kernel ConvNets in capturing global features for image restoration tasks.
- By removing the need for re-parameterizing large convolutional kernels, we have significantly reduced computational consumption. Additionally, we incorporated a simple local feature aggregation network, enabling our model to balance the capture and integration of both global and local features effectively. This innovative approach ensures that our model not only excels in recognizing and enhancing overall image structures but also pays meticulous attention to the finer details, resulting in superior quality enhancements across a variety of document types.
- We developed a clear model composed of simple modules with a minimal number of parameters. Our method is straightforward yet effective, demonstrating that excellent performance can be achieved without the need for complex network architectures. This finding is pivotal for advancing research in the field, as it shows that simplicity can often lead to robust results, encouraging more efficient and accessible approaches to document image enhancement.

2 Related Work

2.1 Document Enhancement

Document enhancement is a crucial field in digital image processing aimed at improving the quality of scanned or captured document images. Key tasks include document deblurring[11], which sharpens images blurred during capture, and document binarization[34], which separates text from the background, enhancing readability for both humans and OCR systems. Other tasks involve document denoising[6], which cleans up visual artifacts like dust and scratches, and shadow removal[36], which corrects brightness inconsistencies caused by shadows across the document. Each of these tasks enhances different aspects of a document's visibility and legibility, making them more suitable for further processing and analysis. Recently, deep learning has significantly contributed to this field, with ConvNets, GANs, and Transformers leading in enhancing document image quality.

Hradiš et al.[10] applied ConvNets to document image deblurring. Souibgui et al.[33] innovatively introduced GANs to the task of document image binarization, developing a framework named DE-GAN aimed at restoring severely degraded document images. They further integrated Vision Transformers (ViT) into the document binarization task, presenting a new encoder-decoder architecture[32] that operates directly on pixel patches to enhance both machine-printed and handwritten document images. Yang et al. introduced DocDiff[37], a document enhancement framework utilizing residual diffusion models to tackle various document enhancement challenges like deblurring and binarization.

As the model performance improves, the computational load increases, and the inference time becomes longer, which poses challenges for practical applications. Document images often have repetitive backgrounds and regular text structures, it is possible to propose a simple yet effective method.

2.2 Large Kernel Convolutional Neural Network

Regarding large-kernel convolutional neural networks, there has been a resurgence of interest due to their appealing performance and efficiency. The transition from small to large kernel convolution marks a significant evolution in convolutional neural network design. Initial explorations like AlexNet[13] utilized larger kernels, but the trend shifted towards smaller kernels due to computational efficiency. However, recent studies have revisited large kernel sizes, finding them advantageous for achieving larger effective receptive fields and better performance on various tasks. For instance, RepLKNet[5] leveraged structural reparameterization to scale kernel size up to 31 × 31, achieving results comparable to or superior to those of Swin Transformer[15].

Furthermore, PeLK[3] proposes a parameter-efficient approach for large kernel ConvNets, introducing peripheral convolution that mimics human vision by efficiently reducing parameter count through parameter sharing, enabling scaling up kernel sizes to 101 × 101 without compromising performance. This method outperformed modern Vision Transformers and ConvNets architectures like Swin[15] and ConvNeXt[16] on several vision tasks, showcasing the potential of large kernels in ConvNets.

3 Method

3.1 Overall Architecture

Fig. 1 shows the overall structure of our method, we opted for a simple U-Net structure. Research by Chen et al.[4] demonstrated that an improved intra-block single-stage U-Net could achieve an optimal balance between computational efficiency and model complexity. Our model employs a four-layer U-Net configuration tailored to a specific structure to ensure efficiency and effectiveness in

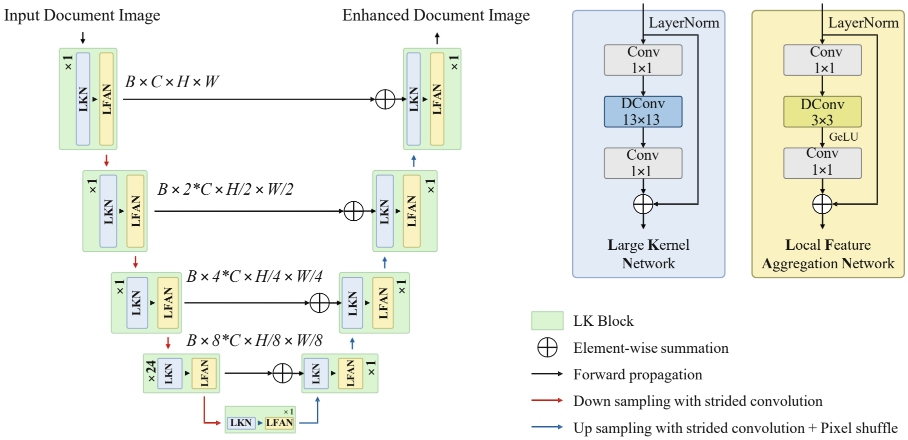


Fig. 1. Overview of LK-Net, the Figure is best readable in 300% zoom.

document image enhancement tasks. In designing the encoder architecture, we referred to RepLKNet, where the network layers are structured as [2, 2, 18, 2]. To reduce computational load in the encoder design, we shifted the largest number of layers to the final stage and reduced the other layers to 1. Since RepLKNet uses 31×31 convolutional layers, to maintain the capability of extracting global features, we increased the number of layers in the final stage to 24. As a result, the encoder comprises a sequence of blocks with the arrangement [1, 1, 1, 24], followed by a middle section consisting of 1 block. To reduce computational load, we kept the decoder with a minimal number of layers. The decoder part inversely mirrors the encoder's complexity, with block numbers set to [1, 1, 1, 1]. This design choice is inspired by advancements in network architecture optimizations, aiming for a balance between model simplicity and performance capability.

Each large kernel block in our network comprises two key modules: the Large Kernel Network (LKN) and the Local Feature Aggregation Network (LFAN). These components are designed to work in tandem, with LKN handling the broad strokes of image processing using large convolutional kernels for global feature extraction and LFAN aggregating the local features to enhance image details and quality.

The backgrounds in document images are usually uniform and repetitive. For example, paper textures, grid lines, or background colors typically remain consistent throughout the image. This consistency makes large convolutional kernels very effective in extracting global background features. Large convolutional kernels can cover larger areas of the image, capturing the global patterns of the background. This helps to remove noise. And the text structures in document images exhibit high repetition and regularity. Whether it's printed text or handwritten text, characters such as letters, numbers, and symbols often have fixed shapes and layouts. This repetition makes small convolutional kernels excel in extracting local text features. Small convolutional kernels can finely capture the edges, curves, and details of the text, ensuring the accurate extraction and restoration of textual information.

Combining large and small kernels forms a powerful multi-scale feature extraction system. Large kernel provide smooth and consistent processing of the global background, while small convolutional kernels focus on detailed text structure extraction and reconstruction. Through this multi-scale feature fusion, the model can better balance global information and local details.

3.2 Large Kernel Network

In Large Kernel Network, we incorporated 13 × 13 depth-wise convolution (DWConv) along with shortcuts to enhance feature extraction capabilities. Following common practices, we employed 1 × 1 convolutions both before and after the DWConv. This approach optimizes the network's computational efficiency and model complexity, effectively enhancing its performance in processing document images.

3.3 Local Feature Aggregation Network

In the Local Feature Aggregation Network, shortcuts are utilized for feature preservation. The network introduces a 1 × 1 convolution to stabilize parameters, followed by a depth-wise (DW) 3 × 3 convolution for local feature extraction. GELU (Gaussian Error Linear Unit)[9] is used to introduce non-linearity, enhancing the network's ability to process complex image textures and details.

3.4 PSNR Loss

We adopted PSNR loss from NAFNet[4] as our loss function for image restoration tasks. Compared to Mean Squared Error (MSE) loss and L1 loss, PSNR loss offers better convergence and outcomes, making it more effective for achieving high-quality image restoration results. Peak Signal-to-Noise Ratio (PSNR) is as follows:

$$\mathrm{PSNR} = 10 \cdot \log\left(\frac{Max^2}{MSE}\right) = \frac{10}{\ln(10)} \cdot \ln(Max^2) - \frac{10}{\ln(10)} \cdot \ln(MSE) \tag{1}$$

By simplifying PSNR, we can derive the PSNR loss as follows:

$$\mathcal{L}_{\mathrm{PSNR}} = \frac{10}{\ln(10)} \cdot \ln(\frac{1}{n}\sum_{i=1}^{n}(pred_i - target_i)^2) \tag{2}$$

4 Experiments and Results

4.1 Document Deblurring

For document deblurring, we conducted our experiments on the widely utilized document deblurring dataset [10], which comprised 66,000 blurred document images with a resolution of 300 × 300 pixels. We randomly selected a subset of this dataset, using 30,000 images for training, 5,000 for validation, and 10,000 for testing.

The evaluation metrics we chose for document deblurring were PSNR and the Structural Similarity Index Measure (SSIM). PSNR is a widely recognized metric that quantifies the quality of reconstructed or processed images by measuring the peak error between the original and the enhanced image. SSIM, on the other hand, evaluated the perceptual quality of the enhanced images by considering the changes in structural information, contrast, and luminance, offering a more comprehensive assessment of image quality that aligns well with human visual perception. We evaluated the computational efficiency of the models by comparing the inference time for processing a single 300 × 300 input image.

As shown in Table 6, our method balanced computational and parameter efficiency, ensuring a lean and powerful network. This efficiency did not come at the cost of performance; on the contrary, in document deblurring, our approach demonstrated exceptional results. Specifically, our method outperformed the current SOTA by a significant margin, achieving a PSNR that was higher by 4.5214 dB and a SSIM that improved by 0.0272. These advancements underscored the effectiveness of our network design in enhancing document image quality, showcasing our model's capability to produce clearer and more accurate images while maintaining a high level of efficiency. Fig. 2 showed that our method achieved better results in detail recovery.

4.2 Document Binarization

For document image binarization, our testing involved the H-DIBCO'14[19], H-DIBCO'18[26], and DIBCO'19[29] datasets. Specifically, for H-DIBCO'14 and H-DIBCO'18, we used the DIBCO'19 dataset as the validation set and a combination of the remaining DIBCO datasets[7,22–25,27,28] (excluding the years under test) and the PALM[2] dataset as the training set. For testing on DIBCO'19, DIBCO'16[27] served as the validation set, while the other DIBCO datasets

Table 1. Result of document deblurring on document deblurring dataset[10].

Method	Parameters	Time	PSNR↑	SSIM↑
DE-GAN[33], TPAMI2020	31M	1.71s	21.5785	0.9029
DocDiff[37], ACM MM2023	**8.20M**	1.68s	23.9818	0.9475
Ours	14.54M	**0.71s**	**28.2149**	**0.9729**

(excluding 2018 and 2019) and PALM formed the training set. Images in the training set were segmented into smaller patches of 256 × 256 pixels.

We aligned our evaluation metrics with those employed in the DIBCO competitions. Specifically, we used PSNR, F-Measure (FM), and F-Measure for pixel-level evaluation (FPS). PSNR served the same purpose as in the other tasks, pro-

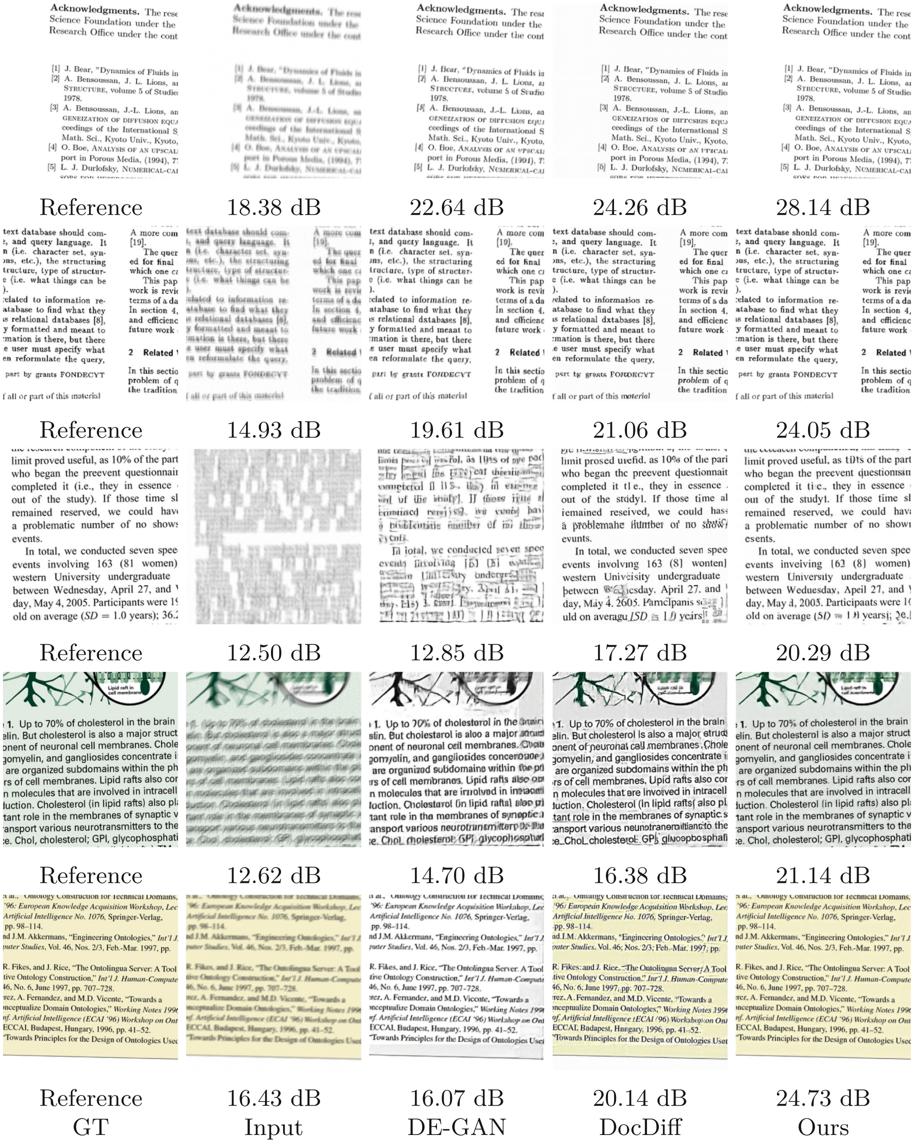

Reference
18.38 dB
22.64 dB
24.26 dB
28.14 dB
Reference
14.93 dB
19.61 dB
21.06 dB
24.05 dB
Reference
12.50 dB
12.85 dB
17.27 dB
20.29 dB
Reference
12.62 dB
14.70 dB
16.38 dB
21.14 dB
Reference
16.43 dB
16.07 dB
20.14 dB
24.73 dB
GT
Input
DE-GAN
DocDiff
Ours

Fig. 2. Qualitative comparison of document deblurring, the Figure is best readable in 300% zoom.

Table 2. Result of document binarization on DIBCO‘19[29] dataset.

Method	Venue	Model	Parameters	Time	DIBCO’19		
					PSNR↑	FM↑	FPS↑
Otsu[20]	TSMC	Thresholding	-	-	9.08	47.83	45.59
Sauvola[31]	PR2000	Thresholding	-	-	13.72	51.73	55.15
Kligler et al.[12]	CVPR2018	CNN	-	-	11.23	53.49	53.34
DE-GAN[33]	TPAMI2020	GAN	31M	40.38s	12.29	55.98	53.44
DocEnTr[32]	ICPR2022	Transformer	68.94M	6.82s	13.43	58.19	53.35
Ours	-	CNN	**14.54M**	**1.90s**	**14.29**	**66.15**	**63.30**

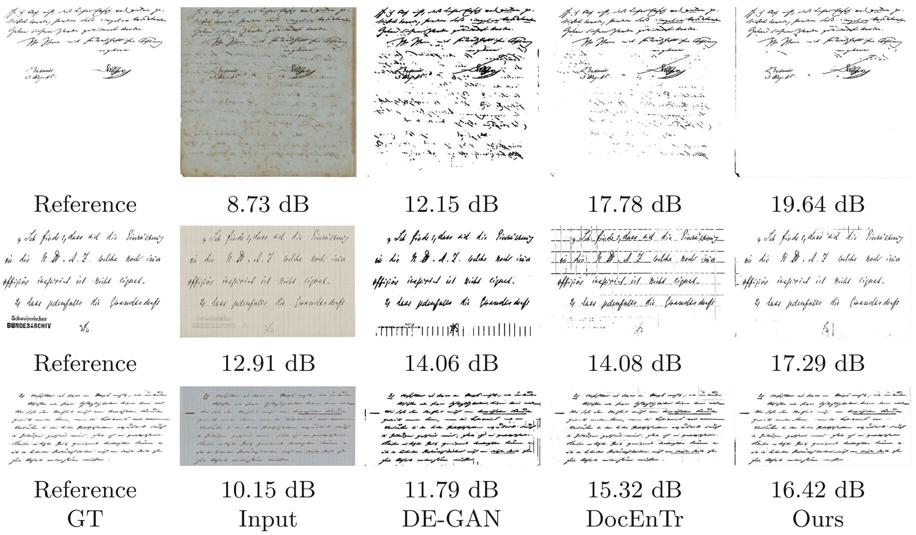


Fig. 3. Qualitative comparison of document binarization on DIBCO’19[29] dataset, the Figure is best readable in 300% zoom.

viding a basis for comparing image quality in terms of error levels. FM assessed the binarization quality by considering both the precision and recall of the binarized text against its ground truth, which was crucial for understanding how well the text was preserved and separated from the background. FPS further refined this assessment by focusing on precision and recall at the pixel level, offering a detailed insight into the model’s performance in accurately rendering the fine details of text and background separation. We choose the largest image in DIBCO‘19 dataset with resolution 2575 × 3465 for comparing the inference time.

As shown in Table 2, our method demonstrated exceptional performance on the DIBCO‘19 dataset. In comparison to the current SOTA, our approach surpassed it with a PSNR improvement of 0.86 dB, an FM increase of 7.96, and a further enhancement in FPS by 9.95. Notably, these significant improvements

Table 3. Result of document binarization on H-DIBCO‘14[19] dataset.

Method	Venue	Model	H-DIBCO’14		
			PSNR↑	FM↑	FPS↑
Otsu[20]	TSMC	Thresholding	18.73	91.62	95.69
Sauvola[31]	PR2000	Thresholding	17.48	83.72	87.49
Zhao et al.[41]	PR2019	GAN	18.37	87.73	90.60
DocEnTr[32]	ICPR2022	Transformer	22.99	97.16	**98.28**
NAFNet[4]	ICCV2022	Transformer	23.29	97.40	97.75
Ours	-	CNN	**24.03**	**97.70**	98.27

were achieved with our model’s parameters being almost one-fifth those of the SOTA model. This achievement not only showcased the efficiency and effectiveness of our method in handling one of the most difficult datasets in document image processing but also highlighted the innovative design of our network that ensured superior performance while maintaining a lean parameter footprint. As demonstrated in Table 3, our method also surpassed SOTA performance in the H-DIBCO‘14 dataset, with a PSNR improvement of 0.74 dB. Specifically, we introduced the natural scenes image restoration model called NAFNet. In terms of document binarization, NAFNet performed well but was slightly inferior to our model. This enhancement further demonstrated our model’s robustness and effectiveness across different benchmarks within the document image processing domain. However, as referred to in Table 4, in the H-DIBCO‘18 dataset, our method achieved performance comparable to, but slightly below, the current SOTA, with a PSNR that was 0.01 dB lower. Fig. 3 demonstrated that our method was more effective in accurately binarizing handwritten text. Although our performance did not surpass the SOTA on DIBCO‘18, Fig. 4 showed that our binarization results were cleaner.

To explore the potential of the model, we also conducted experiments on the GoPro[18] dataset for deblurring in natural scenes. As shown in Table 5, In the task of complex natural scene image restoration, our method is not particularly outstanding. We believe this is due to the low complexity of the model and

Table 4. Result of document binarization on DIBCO‘18[26] dataset.

Method	Venue	Model	DIBCO’18		
			PSNR↑	FM↑	FPS↑
Otsu[20]	TSMC	Thresholding	9.74	51.45	53.05
Sauvola[31]	PR2000	Thresholding	13.78	67.81	74.08
DE-GAN[33]	TPAMI2020	GAN	16.16	77.59	85.74
DocEnTr[32]	ICPR2022	Transformer	**19.47**	**92.53**	**95.15**
Ours	-	CNN	19.46	90.76	94.03

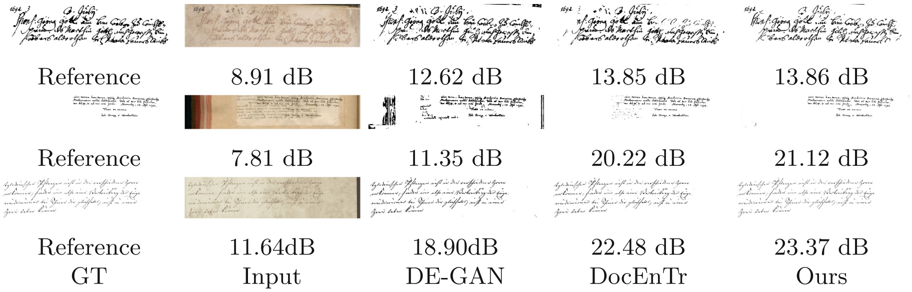


Fig. 4. Qualitative comparison of document binarization on H-DIBCO'18[26] dataset, the Figure is best readable in 300% zoom.

Table 5. Result of natural scene deblurring on GoPro[18] dataset.

Method	Venue	Model	GoPro	
			PSNR↑	SSIM↑
Nah et al.[18]	CVPR2017	CNN	29.08	0.913
DeblurGAN-v2[14]	ICCV2019	GAN	29.55	0.925
Restormer[39]	CVPR2022	Transformer	32.92	0.961
NAFNet[4]	ECCV2022	Transformer	**33.69**	**0.967**
Ours	-	CNN	30.97	0.942

the limited feature extraction capabilities of combining large and small convolution kernels. To achieve better results in natural scenes, it may be necessary to increase the model complexity and the kernel size.

4.3 Ablation Studies

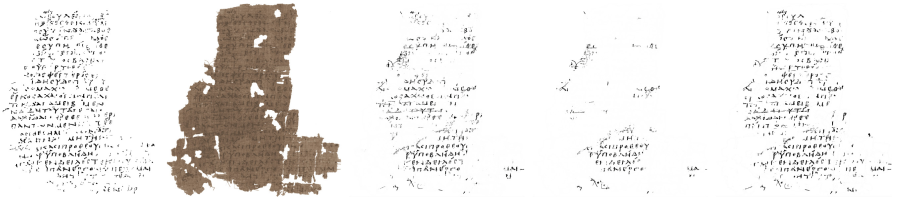
We conducted comparative experiments by changing the convolutional kernels to 3 × 3 and 7 × 7. To ensure a fair comparison, we adjusted the number of blocks in the last encoder to keep the model size similar. The comparison demonstrated that larger kernels had superior performance in document enhancement tasks. This suggested that the architecture benefited significantly from the larger receptive field provided by bigger kernels, enhancing the model's ability to capture and improve document image quality effectively.

As shown in Table 6, our chosen 13 × 13 convolutional kernel yielded better enhancement results. Specifically, PSNR for the 13 × 13 kernel was higher by 0.32 dB compared to the 7 × 7 kernel and by 0.26 dB compared to the 3 × 3 kernel. Moreover, the FM increased by 6.28 over the 7 × 7 kernel and by 3.7 over the 3 × 3 kernel. Similarly, the FPS improved by 5.85 over the 7 × 7 and by 2.93 over the 3 × 3 kernel.

However, Fig. 5 showed that the 7 × 7 kernel did not outperform the 3 × 3 kernel. This indicated that for specific tasks and varying input sizes, selecting

Table 6. Result of document binarization on DIBCO‘19[29] dataset with different kernel size.

Kernel Size	Num of Encoders	Parameters	DIBCO’19		
			PSNR↑	FM↑	FPS↑
3 × 3	[1, 1, 1, 28]	14.97M	14.03	62.45	60.37
7 × 7	[1, 1, 1, 28]	15.31M	13.97	59.87	57.45
13 × 13	[1, 1, 1, 24]	14.54M	**14.29**	**66.15**	**63.30**



Reference GT | 7.56 dB Input | 15.05 dB 3 × 3 | 13.96 dB 7 × 7 | 15.89 dB 13 × 13(Ours)

Fig. 5. Qualitative comparison of document binarization with different kernel size, the Figure is best readable in 300% zoom.

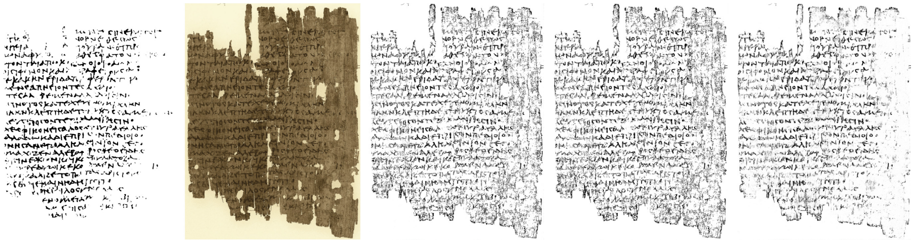
the appropriate size of the large convolutional kernel was crucial to achieving optimal results. If the size of the convolutional kernel chosen was neither large nor small enough, it was likely that it would neither capture global features well nor local features effectively. This highlighted the importance of tailoring the kernel size according to the specific requirements of the task at hand, as larger kernels could capture wider contextual information but might not always be the most efficient choice depending on the complexity and characteristics of the input data.

To validate the effectiveness of PSNR loss, we conducted additional training on the DIBCO‘19 dataset using MSE loss for comparison. Fig. 6 showed that PSNR loss achieved better outcomes and it had a better effect on removing the background of the text.

In typical image restoration tasks, researchers often used MSE loss and L1 loss as their primary loss functions. However, as illustrated in Table 7, using PSNR as a loss function in our experiments yielded more favorable results compared to using MSE and L1 loss. Specifically, the PSNR was higher by 0.42 dB than when using MSE and L1 loss. Furthermore, the FM improved by 1.3 over MSE and by 1.46 over L1 loss. In terms of FPS, there was an increase of 0.32 over MSE and 0.42 over L1 loss. This observation underscored the potential of PSNR loss in the field of image restoration, suggesting that it may be a more effective measure for optimizing image quality during the restoration process.

Table 7. Result of document binarization on DIBCO'19[29] dataset with different loss functions.

Loss	DIBCO'19		
	PSNR↑	FM↑	FPS↑
L1	13.87	64.69	62.88
MSE	13.87	64.85	62.98
PSNR	**14.29**	**66.15**	**63.30**



Reference GT

5.65 dB Input

8.76 dB L1

8.66 dB MSE

10.08 dB PSNR(Ours)

Fig. 6. Qualitative comparison of document binarization with different loss functions, the Figure is best readable in 300% zoom.

5 Conclusions

In this paper, we introduce a simple and novel large kernel network, achieving SOTA performance across multiple document enhancement benchmarks. We constructed a baseline for document enhancement using a convolutional neural network, incorporating shortcuts and the GELU activation function. We believe that this approach will facilitate future research by providing a straightforward benchmark for comparative experiments. This baseline setup is designed to be both effective and easy to replicate, offering a solid foundation for further innovation and evaluation in the field of document image processing. This approach underscores the effectiveness of large convolutional kernels in improving document image quality. We believe this work will redirect researchers' attention towards designing simple and effective networks to enhance the performance of document enhancement.

Acknowledgements. This work was supported by the National Natural Science Foundation of China under Grant No.62176091.

References

1. Awal, A.M., Ghanmi, N., Sicre, R., Furon, T.: Complex document classification and localization application on identity document images. In: 2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR). vol. 01, pp. 426–431 (2017). https://doi.org/10.1109/ICDAR.2017.77

2. Burie, J.C., Coustaty, M., Hadi, S., Kesiman, M.W.A., Ogier, J.M., Paulus, E., Sok, K., Sunarya, I.M.G., Valy, D.: Icfhr2016 competition on the analysis of handwritten text in images of balinese palm leaf manuscripts. In: 2016 15th International Conference on Frontiers in Handwriting Recognition (ICFHR). pp. 596–601 (2016https://doi.org/10.1109/ICFHR.2016.0114
3. Chen, H., Chu, X., Ren, Y., Zhao, X., Huang, K.: Pelk: Parameter-efficient large kernel convnets with peripheral convolution. arXiv preprint arXiv:2403.07589 (2024)
4. Chen, L., Chu, X., Zhang, X., Sun, J.: Simple baselines for image restoration. arXiv preprint arXiv:2204.04676 (2022)
5. Ding, X., Zhang, X., Zhou, Y., Han, J., Ding, G., Sun, J.: Scaling up your kernels to 31x31: Revisiting large kernel design in cnns. arXiv preprint arXiv:2203.06717 (2022)
6. Fan, L., Fan, L., Tan, C.: Wavelet diffusion for document image denoising. In: Seventh International Conference on Document Analysis and Recognition, 2003. Proceedings. pp. 1188–1192 (2003https://doi.org/10.1109/ICDAR.2003.1227845
7. Gatos, B., Ntirogiannis, K., Pratikakis, I.: Icdar 2009 document image binarization contest (dibco 2009). In: 2009 10th International Conference on Document Analysis and Recognition. pp. 1375–1382 (2009).https://doi.org/10.1109/ICDAR.2009.246
8. Goodfellow, I.J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. In: Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2. p. 2672-2680. NIPS'14, MIT Press, Cambridge, MA, USA (2014)
9. Hendrycks, D., Gimpel, K.: Gaussian error linear units (gelus). arXiv: Learning (2016), https://api.semanticscholar.org/CorpusID:125617073
10. Hradiš, M., Kotera, J., Zemčík, P., Šroubek, F.: Convolutional neural networks for direct text deblurring. In: British Machine Vision Conference (2015), https://api.semanticscholar.org/CorpusID:14143575
11. Jiao, J., Sun, J., Satoshi, N.: A convolutional neural network based two-stage document deblurring. In: 2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR). vol. 01, pp. 703–707 (2017).https://doi.org/10.1109/ICDAR.2017.120
12. Kligler, N., Katz, S., Tal, A.: Document enhancement using visibility detection. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2374–2382 (2018https://doi.org/10.1109/CVPR.2018.00252
13. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: Proceedings of the 25th International Conference on Neural Information Processing Systems - Volume 1. p. 1097-1105. NIPS'12, Curran Associates Inc., Red Hook, NY, USA (2012)
14. Kupyn, O., Martyniuk, T., Wu, J., Wang, Z.: Deblurgan-v2: Deblurring (orders-of-magnitude) faster and better. In: The IEEE International Conference on Computer Vision (ICCV) (Oct 2019)
15. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) (2021)
16. Liu, Z., Mao, H., Wu, C.Y., Feichtenhofer, C., Darrell, T., Xie, S.: A convnet for the 2020s. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2022)
17. Naeem, M.F., Zia, N.u.S., Awan, A.A., Shafait, F., ul Hasan, A.: Impact of ligature coverage on training practical urdu ocr systems. In: 2017 14th IAPR International

Conference on Document Analysis and Recognition (ICDAR). vol. 01, pp. 131–136 (2017https://doi.org/10.1109/ICDAR.2017.30

18. Nah, S., Kim, T.H., Lee, K.M.: Deep multi-scale convolutional neural network for dynamic scene deblurring (2018), https://arxiv.org/abs/1612.02177
19. Ntirogiannis, K., Gatos, B., Pratikakis, I.: Icfhr2014 competition on handwritten document image binarization (h-dibco 2014). In: 2014 14th International Conference on Frontiers in Handwriting Recognition. pp. 809–813 (2014https://doi.org/10.1109/ICFHR.2014.141
20. Otsu, N.: A threshold selection method from gray-level histograms. IEEE Trans. Syst. Man Cybern. **9**(1), 62–66 (1979). https://doi.org/10.1109/TSMC.1979.4310076
21. Parker, J., Frieder, O., Frieder, G.: Automatic enhancement and binarization of degraded document images. In: 2013 12th International Conference on Document Analysis and Recognition. pp. 210–214 (2013https://doi.org/10.1109/ICDAR.2013.49
22. Pratikakis, I., Gatos, B., Ntirogiannis, K.: H-dibco 2010 - handwritten document image binarization competition. In: 2010 12th International Conference on Frontiers in Handwriting Recognition. pp. 727–732 (2010).https://doi.org/10.1109/ICFHR.2010.118
23. Pratikakis, I., Gatos, B., Ntirogiannis, K.: Icdar 2011 document image binarization contest (dibco 2011). In: Proceedings of the 2011 International Conference on Document Analysis and Recognition. p. 1506-1510. ICDAR '11, IEEE Computer Society, USA (2011https://doi.org/10.1109/ICDAR.2011.299, https://doi.org/10.1109/ICDAR.2011.299
24. Pratikakis, I., Gatos, B., Ntirogiannis, K.: Icfhr 2012 competition on handwritten document image binarization (h-dibco 2012). In: 2012 International Conference on Frontiers in Handwriting Recognition. pp. 817–822 (2012https://doi.org/10.1109/ICFHR.2012.216
25. Pratikakis, I., Gatos, B., Ntirogiannis, K.: Icdar 2013 document image binarization contest (dibco 2013). In: 2013 12th International Conference on Document Analysis and Recognition. pp. 1471–1476 (2013).https://doi.org/10.1109/ICDAR.2013.219
26. Pratikakis, I., Zagori, K., Kaddas, P., Gatos, B.: Icfhr 2018 competition on handwritten document image binarization (h-dibco 2018). In: 2018 16th International Conference on Frontiers in Handwriting Recognition (ICFHR). pp. 489–493 (2018https://doi.org/10.1109/ICFHR-2018.2018.00091
27. Pratikakis, I., Zagoris, K., Barlas, G., Gatos, B.: Icfhr2016 handwritten document image binarization contest (h-dibco 2016). In: 2016 15th International Conference on Frontiers in Handwriting Recognition (ICFHR). pp. 619–623 (2016https://doi.org/10.1109/ICFHR.2016.0118
28. Pratikakis, I., Zagoris, K., Barlas, G., Gatos, B.: Icdar2017 competition on document image binarization (dibco 2017). In: 2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR). vol. 01, pp. 1395–1403 (2017https://doi.org/10.1109/ICDAR.2017.228
29. Pratikakis, I., Zagoris, K., Karagiannis, X., Tsochatzidis, L., Mondal, T., Marthot-Santaniello, I.: Icdar 2019 competition on document image binarization (dibco 2019). In: 2019 International Conference on Document Analysis and Recognition (ICDAR). pp. 1547–1556 (2019).https://doi.org/10.1109/ICDAR.2019.00249
30. Santos, J.E.B.d.: Automatic content extraction on semi-structured documents. In: 2011 International Conference on Document Analysis and Recognition. pp. 1235–1239 (2011https://doi.org/10.1109/ICDAR.2011.249

31. Sauvola, J., Seppanen, T., Haapakoski, S., Pietikainen, M.: Adaptive document binarization. In: Proceedings of the Fourth International Conference on Document Analysis and Recognition. vol. 1, pp. 147–152 vol.1 (1997).https://doi.org/10.1109/ICDAR.1997.619831
32. Souibgui, M.A., Biswas, S., Jemni, S.K., Kessentini, Y., Fornés, A., Lladós, J., Pal, U.: Docentr: An end-to-end document image enhancement transformer. In: 2022 26th International Conference on Pattern Recognition (ICPR) (2022)
33. Souibgui, M.A., Kessentini, Y.: De-gan: A conditional generative adversarial network for document enhancement. IEEE Trans. Pattern Anal. Mach. Intell. (2020). https://doi.org/10.1109/TPAMI.2020.3022406
34. Tensmeyer, C., Martinez, T.: Document image binarization with fully convolutional neural networks. In: 2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR). vol. 01, pp. 99–104 (2017).https://doi.org/10.1109/ICDAR.2017.25
35. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention is all you need. In: Proceedings of the 31st International Conference on Neural Information Processing Systems. p. 6000-6010. NIPS'17, Curran Associates Inc., Red Hook, NY, USA (2017)
36. Wang, J.R., Chuang, Y.Y.: Shadow removal of text document images by estimating local and global background colors. In: ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 1534–1538 (2020https://doi.org/10.1109/ICASSP40776.2020.9053378
37. Yang, Z., Liu, B., Xxiong, Y., Yi, L., Wu, G., Tang, X., Liu, Z., Zhou, J., Zhang, X.: Docdiff: Document enhancement via residual diffusion models. In: Proceedings of the 31st ACM International Conference on Multimedia. pp. 2795–2806 (2023)
38. Zagoris, K., Pratikakis, I.: Bio-inspired modeling for the enhancement of historical handwritten documents. In: 2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR). vol. 01, pp. 287–292 (2017).https://doi.org/10.1109/ICDAR.2017.55
39. Zamir, S.W., Arora, A., Khan, S., Hayat, M., Khan, F.S., Yang, M.H.: Restormer: Efficient transformer for high-resolution image restoration. In: CVPR (2022)
40. Zhai, X., Kolesnikov, A., Houlsby, N., Beyer, L.: Scaling vision transformers. 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) pp. 1204–1213 (2021), https://api.semanticscholar.org/CorpusID:235367962
41. Zhao, J., Shi, C., Jia, F., Wang, Y., Xiao, B.: Document image binarization with cascaded generators of conditional generative adversarial networks. Pattern Recogn. **96**(C) (dec 2019).https://doi.org/10.1016/j.patcog.2019.106968, https://doi.org/10.1016/j.patcog.2019.106968

ML-CrAIST: Multi-scale Low-High Frequency Information-Based Cross Attention with Image Super-Resolving Transformer

Alik Pramanick(✉), Utsav Bheda, and Arijit Sur

Indian Institute of Technology Guwahati, Guwahati, India
{p.alik,u.bheda,arijit}@iitg.ac.in

Abstract. Recently, transformers have captured significant interest in the area of single-image super-resolution tasks, demonstrating substantial gains in performance. Current models heavily depend on the network's extensive ability to extract high-level semantic details from images while overlooking the effective utilization of multi-scale image details and intermediate information within the network. Furthermore, it has been observed that high-frequency areas in images present significant complexity for super-resolution compared to low-frequency areas. This work proposes a transformer-based super-resolution architecture called ML-CrAIST that addresses this gap by utilizing low-high frequency information in multiple scales. Unlike most of the previous work (either spatial or channel), we operate spatial and channel self-attention, which concurrently model pixel interaction from both spatial and channel dimensions, exploiting the inherent correlations across spatial and channel axis. Further, we devise a cross-attention block for super-resolution, which explores the correlations between low and high-frequency information. Quantitative and qualitative assessments indicate that our proposed ML-CrAIST surpasses state-of-the-art super-resolution methods (e.g., 0.15 dB gain @Manga109 ×4). Code is available at: https://github.com/Alik033/ML-CrAIST.

Keywords: Transformer · Image super-resolution · Spatial domain · Frequency domain · Cross attention

1 Introduction

The task of single image super-resolution (SR) [7] remains an enduring low-level challenge that centers on the restoration of high-resolution (HR) images from degraded low-resolution (LR) inputs. As an issue with inherent ambiguity and numerous possible solutions for a given LR image, several methods have emerged in recent years to address and overcome this challenge. Numerous methods in this context use convolution neural networks (CNNs) [9,10,15,32,35,47] to improve performance in a variety of applications. These methods mostly use

A. Antonacopoulos et al. (Eds.): ICPR 2024, LNCS 15321, pp. 291–307, 2025.
https://doi.org/10.1007/978-3-031-78305-0_19

residual learning [15], dense connections [35], or channel attention [47] to build network architectures, significantly contributing to developing super-resolution models. However, the CNN-based approach exhibits a limited receptive field due to the localized nature of convolution, which hampers the global dependencies, consequently restricting the overall performance of the model.

In recent times, the Transformer architecture, initially introduced in natural language processing (NLP), has demonstrated significant success across a wide range of high-level vision tasks [3,4,40]. This success is attributed to its incorporation of a self-attention mechanism, which effectively establishes global dependencies. A notable advancement in SR is SwinIR [20], which presents the Swin Transformer, leading to significant enhancements over state-of-the-art CNN-based models across different standard datasets. Subsequent developments, including Swin-FIR [43], ELAN [45], and HAT [6], have extended the capabilities of SwinIR by utilizing Transformers to develop various network architectures for SR tasks. These methods demonstrate that appropriately enlarging the windows for the shifted window self-attention in SwinIR can lead to obvious improvements in performance. However, the increase in computational burden becomes a significant concern as the window size grows more prominent. Furthermore, Transformer-based methods rely on self-attention and need networks with more channels than previous CNN-based methods [1,14,16]. Also, they use uni-dimensional aggregation operations (either spatial or channel) and homogeneous aggregation schemes (simple hierarchical stacking of convolution and self-attention). Wang et al. [37] consider the above problem and design OmniSR to achieve superior performance. Despite substantial progress in super-resolution methods, they even encounter visual artifacts in the resulting images, such as inadequate texture representation and loss of details. Further, it has been observed that super-resolving high-frequency image areas are more challenging than low-frequency areas. Numerous existing SR methods work solely within the spatial domain, concentrating on improving the resolution of low-resolution pixels to obtain a high-resolution image. They often overlook the potential benefits of the frequency domain, which could offer a better method for retrieving lost high-frequency information. Also, it needs to include more texture patterns of multi-scales, which is required in SR tasks. Similar textures with multiple scales may exist within a single image at different positions. For instance, repetitive patterns at different scales (such as facades, windows, etc., in a building) may appear in various locations within a single image. The multi-scale aware framework is required to use the beneficial non-local detail, which aggregates the information from all the different scales of the LR image.

To address the above mentioned issues and achieve higher performance, this work proposes a novel super-resolution model that simultaneously exploits frequency and spatial domain information at different scales. 2D Discrete Wavelet Transformation (2dDWT) is used to analyze both the high (LH, HL, and HH) and low (LL) frequency wavelet sub-bands. To carefully design a cross-attention block, we fuse low and high frequency information to boost SR performance. We explore the features in multiple scales and systematically combine information across all scales at each resolution level, facilitating meaningful information exchange. Simultaneously, another fusion technique is proposed to combine the high-frequency sub-bands while maintaining their unique complementary characteristics that differ from simple concatenation or averaging of the sub-bands. The major contributions of this paper are as follows:

1. A novel multi-scale model is proposed by utilizing both spatial and frequency domain features that is capable to enhance the spatial resolution of an low-resolution image.
2. In addition, a low-high frequency interaction block (LHFIB) is introduced to exchange the information between low and high frequency sub-bands through the proposed cross attention block (CAB).
3. A non-linear approach is proposed to fuse high-frequency sub-bands using an attention mechanism for more precise restoration of high-frequency details.
4. Informative features are obtained from different scales using CAB while preserving the high-resolution features to represent spatial details accurately.

2 Related Work

- **Conventional CNNs for SR.** CNNs have achieved remarkable success in the task of image super-resolution. SRCNN [10] is notable as the pioneering CNN-based super-resolution method, outperforming the performance of traditional approaches (e.g., bicubic, nearest-neighbor, and bilinear interpolation). After this initial advancement, significant attention has been directed towards expanding the network depth and incorporating residual learning techniques to enhance super-resolution performance [15,35,47]. EDSR [21] further improves peak signal-to-noise ratio (PSNR) results significantly by removing the unnecessary Batch Normalization layers. Additionally, RCAN [47] integrates a channel attention mechanism to enhance feature aggregation efficiency, enabling improved performance even with deeper network architectures. Subsequent models such as SAN [9], NLSA [28], and HAN [29] have introduced a range of attention mechanisms, either focusing on spatial or channel dimensions, reflecting a growing trend in attention-based approaches within the field. To improve reconstruction quality while working within constrained computing resources, DRCN [16], DRRN [34], CARN [1], IMDN [14] delve into lightweight architectural designs. Another research direction is operating model compression strategies like knowledge distillation [11,46] and neural architecture search [8] to decrease computing costs.
- **Generative adversarial networks (GANs) for SR.** GANs [12] provide a fundamental method to balance perception and distortion by regulating the weights of perceptual and fidelity losses, generating realistic images. [18] introduced SRGAN, which incorporates adversarial training with the SRResNet generator. [38] presented ESRGAN featuring the residual-in-residual dense block framework for super-resolution. Later, [33] enhanced ESRGAN by auxiliary noise injection and proposed ESRGAN+. Park et

al. [31] suggested Flexible Style SR, which optimizes the SR model with image specific objectives without viewing the regional features. These methods [18,30,33,38] suffer from the computational burden posed by numerous image maps.

■ **Transformer-based methods for SR.** Recently, Transformers have shown significant promise in a range of vision tasks, including image classification [40], object detection [4], semantic segmentation [3], image restoration [5,22,39], etc. Among these approaches, the most prominent example is the Vision Transformer (ViT), demonstrating that transformers can outperform convolutional neural networks in feature encoding tasks. Designing transformer-based models for image super-resolution poses a significant challenge as it requires preserving the structural details of the input image. IPT [5] is a pre-trained model built upon the transformer encoder and decoder structure and has been used for super-resolution. SwinIR [20] employs a window-based attention mechanism to tackle image super-resolution tasks, demonstrating superior performance over CNN-based methods. ELAN [45] facilitates the architecture of SwinIR and utilizes self-attention calculated in different window sizes to capture correlations between long-range pixels. Choi et al. [7] introduce N-gram context into low-level vision tasks using Transformers for the SR task. Most recently, OmniSR [37] explored spatial-channel axis aggregation networks to enhance SR performance.

Our approach also relies on the transformer architecture. Unlike the aforementioned methods, which predominantly utilize spatial domain information and compute self-attention for model construction, our primary focus is on leveraging spatial-frequency domain features and multi-scale features through cross-attention to improve the performance of the super-resolution model.

3 Proposed Method

Figure 1 shows the proposed architecture that aims to generate an SR image from the degraded LR image.

3.1 Overall Pipeline

This section presents a comprehensive description of the overall network architecture. Given an LR image $I_{LR} \in \mathbb{R}^{H \times W \times 3}$, we pass it through a convolution layer to extract the initial feature f_0. The acquired feature is then fed into N spatial-channel attention-based transformer blocks (SCATB), from which the deep spatial and channel-wise correlated features f_d are extracted.

$$f_0 = \mathcal{C}^{3\times 3}(I_{LR}), \quad f_i = \mathcal{F}^i_{\mathcal{SCATB}}(f^{i-1}), \quad f_d = f_N, \tag{1}$$

where $\mathcal{C}^{3\times 3}$ refers a convolution with 3×3 kernel size, $\mathcal{F}^i_{\mathcal{SCATB}}$ represents the i-th SCATB, and $f_1, f_2, .., f_N$ denote intermediate features.

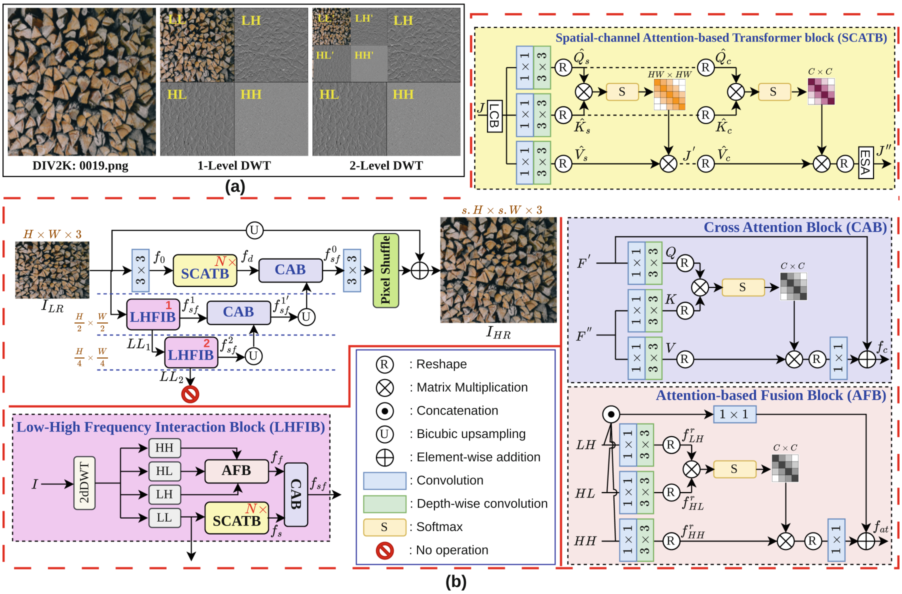


Fig. 1. (a) Multi-level wavelet sub-bands of a LR image. (b) Overview of the Proposed ML-CrAIST. $N\times$ indicates that the block is stacked N times.

Simultaneously, we input I_{LR} into the first low-high frequency interaction block (LHFIB) to extract spatial-frequency information $f^1_{sf} \in \mathbb{R}^{\frac{H}{2}\times\frac{W}{2}\times c}$ and LL cube. The LL cube of the first LHFIB is fed into the second LHFIB to extract further spatial-frequency information ($f^2_{sf} \in \mathbb{R}^{\frac{H}{4}\times\frac{W}{4}\times c}$) in different scales. Each LHFIB contains an attention-based fusion block (AFB) to fuse the high-frequency sub-bands, N number of SCATBs to capture spatially and channel-wise refined features from the low-frequency sub-band, and a cross attention block (CAB) for message passing between refined low-high frequency features. Next, we up-sample f^2_{sf} and combine it with f^1_{sf} within the cross-attention block (CAB) to obtain informative multi-scale features, denoted as $f^{1'}_{sf} \in \mathbb{R}^{\frac{H}{2}\times\frac{W}{2}\times c}$. Next, we up-sample the $f^{1'}_{sf}$ feature and fed it alongside f_d into the CAB module to generate meaningful features ($f^0_{sf} \in \mathbb{R}^{H\times W\times c}$) that contain refined multi-scales feature information.

$$\begin{aligned} f^1_{sf}, LL_1 = \mathcal{F}_{LHFIB}(I_{LR}), \quad f^2_{sf}, LL_2 = \mathcal{F}_{LHFIB}(LL_1), \\ f^{1'}_{sf} = \mathcal{F}_{CAB}(f^1_{sf}, \mathcal{U}(f^2_{sf})), \quad f^0_{sf} = \mathcal{F}_{CAB}(f_d, \mathcal{U}(f^1_{sf})), \end{aligned} \tag{2}$$

where, $\mathcal{F}_{LHFIB}$, $\mathcal{F}_{CAB}$, and $\mathcal{U}$ represent the LHFIB, CAB and bicubic up-sampling operation. Next, we employ a convolution layer and set the output channels to $3s^2$, where s denotes the scale factor by which the spatial resolution is to be enhanced. Finally, a PixelShuffle ($\mathcal{PS}$) layer takes the low-resolution feature maps ($f_l \in \mathbb{R}^{H\times W\times 3s^2}$) and produce the high-resolution image

$I_{HR} \in \mathbb{R}^{s.H \times s.W \times 3}$. Then, the reconstructed HR image I_{HR} can be written as

$$I_{HR} = \mathcal{PS}(f_l) + \mathcal{U}(I_{LR}), \quad f_l = \mathcal{C}^{3\times 3}(f_{sf}^0) \tag{3}$$

The proposed ML-CrAIST is optimized using the $\mathcal{L}_1$ loss:

$$\mathcal{L}_1(I_{HR}^g, I_{HR}) = \frac{1}{M}\sum_{a=1}^{M} \|(I_{HR})^a - (I_{HR}^g)^a\|_1, \tag{4}$$

where I_{HR}^g indicates the ground-truth image.

3.2 Spatial-channel attention-based transformer block (SCATB)

Wang et al. [37] introduced the omni-self attention (OSA) block, which has been integrated to capture pixel interactions from spatial and channel dimensions simultaneously, enabling the exploration of potential correlations across spatial and channel dimensions. Instead of using a standard transformer block, we leverage the OSA block along with LCB [37] and ESA [17] block as a SCATB to capture useful local details and long-range dependencies effectively.

To formally define its operational principle, let $J \in \mathbb{R}^{H\times W\times C}$ be the intermediate feature map that passes through an LCB block ($\mathcal{F}_{LCB}$) to aggregate local contextual information (f_c^l), then SCATB generates query (Q), key (K) and value (V) projections by using a 1×1 convolution ($\mathcal{C}^{1\times 1}$) followed by 3×3 depth-wise convolution ($\mathcal{D}_c^{3\times 3}$) on f_c^l, where $Q, K, V \in \mathbb{R}^{H\times W\times C}$. Next, we reshaped query ($\hat{Q_s} \in \mathbb{R}^{HW\times C}$), key ($\hat{K_s} \in \mathbb{R}^{C\times HW}$), and value ($\hat{V_s} \in \mathbb{R}^{HW\times C}$) projections, and calculate the attention map of size $\mathbb{R}^{HW\times HW}$ between $\hat{Q_s}$ and $\hat{K_s}$ in spatial dimension which is multiplied with the $\hat{V_s}$ to get the spatially enriched attentive features $J^{'}$. Next stage, to get the attention map of size $\mathbb{R}^{C\times C}$ in channel dimension, we reshape query ($\hat{Q_c} \in \mathbb{R}^{C\times HW}$), key ($\hat{K_c} \in \mathbb{R}^{HW\times C}$) and $J^{'}$ as value ($\hat{V_c} \in \mathbb{R}^{C\times HW}$). Then, we perform the matrix multiplication between $\hat{Q_c}$ and $\hat{K_c}$ followed by a softmax operation and get the channel-wise attentive feature map. Finally, the channel-wise attentive feature maps are multiplied with the $\hat{V_c}$ and get the spatial and channel-wise correlated feature maps. Lastly, these feature maps are fed into the ESA block ($\mathcal{F}_{ESA}$) to refine the features further. Overall, the procedure is described as:

$$\begin{aligned} Q, K, V = \mathcal{D}_c^{3\times 3}(\mathcal{C}^{1\times 1}(f_c^l)), \quad f_c^l = \mathcal{F}_{LCB}(J), \quad \hat{K_s} = \mathcal{R}(K), \\ \hat{Q_s} = \mathcal{R}(Q), \quad \hat{V_s} = \mathcal{R}(V), \quad J^{'} = \mathcal{S}(\hat{K_s}.\hat{Q_s}).\hat{V_s}, \quad \hat{K_c} = \mathcal{R}(\hat{K_s}), \\ \hat{Q_c} = \mathcal{R}(\hat{Q_s}), \quad \hat{V_c} = \mathcal{R}(J^{'}), \quad J^{''} = \mathcal{F}_{SCATB}(J) = \mathcal{F}_{ESA}(\mathcal{S}(\hat{K_c}.\hat{Q_c}).\hat{V_c}), \end{aligned} \tag{5}$$

where $\mathcal{S}$, $\mathcal{R}$, and $\mathcal{F}_{SCATB}$, indicate the softmax function, reshape, and spatial-channel attention-based transformer operation, respectively. We encourage the reader to refer [37] for more details. We have demonstrated that OSA is advantageous over standard transformer block [41] in producing visually pleasing SR images in the experiments section.

3.3 Low-High Frequency Interaction Block (LHFIB)

In this work, to integrate frequency domain information with spatial domain, we apply the Haar wavelet transformation as a 2D discrete wavelet transformation to the LR image (I_{LR}) and decompose it into four sub-bands (LL, LH, HL, and HH) where every sub-band $\in \mathbb{R}^{\frac{H}{2} \times \frac{W}{2} \times 1}$. The LL sub-band characterizes the background details within the image, while LH, HL, and HH sub-bands characterize variations along vertical axis, variations along horizontal axis, and diagonal information present in the image. The LL sub-band and the original degraded image are typically employed for analyzing spatial information. Since LH, HL, and HH sub-bands preserve high-frequency components, they provide richer content for enhancing high-frequency detail during the super-resolution process. To leverage the benefit of frequency and spatial details, we design a low-high frequency interaction block.

In detail, let it take I as input and break it down into LL, LH, HL, and HH components. Next, we combine the high-frequency sub-bands (i.e., LH, HL, and HH) using an attention-based fusion block (AFB) and get the refined high-frequency information f_f. The low-frequency (i.e., LL) sub-band is fed into SCATB to extract useful spatial information f_s. Finally, we have performed the cross-attention between low and high-frequency features to enable intelligent feature aggregation. The entire approach can be formulated as:

$$\begin{aligned} LL, LH, HL, HH = \mathcal{F}_{DWT}(I), \quad f_f = \mathcal{F}_{AFB}(LH, HL, HH), \\ f_s = \mathcal{F}_{SCATB}(LL), \quad f_{sf} = \mathcal{F}_{CAB}(f_f, f_s), \quad f_{sf}, LL = \mathcal{F}_{LHFIB}(I), \end{aligned} \tag{6}$$

where $\mathcal{F}_{DWT}$, $\mathcal{F}_{AFB}$, $\mathcal{F}_{CAB}$ and $\mathcal{F}_{LHFIB}$ refer 2dDWT, attention-based fusion, cross-attention, and low-high frequency interaction operation, respectively.

3.4 Attention-based fusion block (AFB)

The conventional method for feature aggregation typically involves either simple concatenation or summation. However, these types of selection offer restricted expressive capabilities of the network, as [19] suggested. In this context, we present a nonlinear method for merging features through an attention mechanism to identify and amplify the more relevant features. As shown in Figure 1, we propose an attention-based fusion block (AFB) to combine the high-frequency cubes so that only useful information can be processed further. We pass the high-frequency sub-bands through a convolution layer with 1 × 1 kernel size and a depth-wise convolution layer with 3 × 3 kernel size. Next, we reshape the features to obtain $f_{LL}^r, f_{HH}^r \in \mathbb{R}^{C \times HW}$ and $f_{HL}^r \in \mathbb{R}^{HW \times C}$. We compute the matrix multiplication between f_{LH}^r and f_{HL}^r followed by a softmax operation to get the attentive map (f_a) of size $\mathbb{R}^{C \times C}$. This attention map f_a is multiplied with f_{HH}^r to obtain attentive feature f_{at}. Finally, the concatenated LH, HL, and HH sub-bands are convolved through a 1 × 1 convolution and added with the reshaped attentive feature to produce the attention-based fused high-frequency features. Such an operation can be defined as:

$$\begin{aligned} f_{sb} = \mathcal{D}_c^{3\times3}(\mathcal{C}^{1\times1}(sb)), \quad f_{sb}^r = \mathcal{R}(f_{sb}), \quad sb \in \{LH, HL, HH\} \\ f_{at} = \mathcal{F}_{AFB}(LH, HL, HH) = \mathcal{C}^{1\times1}(LH \odot HL \odot HH) \\ +\mathcal{C}^{1\times1}(\mathcal{R}(\mathcal{S}(f_{LH}^r . f_{HL}^r) . f_{HH}^r)), \end{aligned} \tag{7}$$

where $\mathcal{S}$, $\mathcal{R}$, $\odot$ refer to softmax function, reshape operation, and concatenation operation, respectively. Through ablation, we have shown that the AFB yields more promising outcomes than regular concatenation and addition.

3.5 Cross Attention Block (CAB)

CAB integrates two distinct embedding sequences of identical dimensions. It employs query from one sequence and key and value from the other. The attention masks from one embedding sequence are used to emphasize the extracted features in another embedding sequence. We introduce two cross-attention blocks (CAB) with similar architectures for message passing: one operates between low-high frequency features, and the other operates between multi-scale features. For low-high frequency features, it leverages the low frequency features $(F^{'})$ to generate a query projection and employs high frequency features $(F^{''})$ to create key and value projections through a standard 1×1 convolution and a 3×3 depth-wise convolution layer. Similarly, in the multi-scale scenario, one scale feature $(F^{'})$ is used to generate the query projection, while another scale feature $(F^{''})$ is used to create the key and value projections. Overall, cross-attention can be obtained by

$$\begin{aligned} Q = \mathcal{D}_c^{3\times3}(\mathcal{C}^{1\times1}(F^{'})), \quad K, V = \mathcal{D}_c^{3\times3}(\mathcal{C}^{1\times1}(F^{''})), \quad Q_r = \mathcal{R}(Q), \\ K_r = \mathcal{R}(K), \quad V_r = \mathcal{R}(V), \quad \mathcal{CA}(Q_r, K_r, V_r) = \mathcal{S}(Q_r \cdot K_r) \cdot V_r, \\ f_c = \mathcal{F}_{\mathcal{CAB}}(Q, K, V) = \mathcal{C}^{1\times1}(\mathcal{R}(\mathcal{CA}(Q_r, K_r, V_r))) + F^{'}, \end{aligned} \tag{8}$$

where $Q, V \in \mathbb{R}^{C\times HW}$, $K \in \mathbb{R}^{HW\times C}$, and $\mathcal{CA}$ represents the cross-attention function.

4 Experiments

4.1 Datasets & Evaluation Metrics

Following prior research [7,20,37], we employ the DIV2K dataset [36] for training. For testing purposes, we utilize five widely recognized benchmark datasets: Set5 [2], Set14 [42], B100 [26], Urban100 [13], and Manga109 [27]. The testing results are assessed based on PSNR and structural similarity index measure (SSIM) values computed on the Y channel (i.e., luminance) within the YCbCr color space. Also, we evaluate the learned perceptual image patch similarity (LPIPS) metrics. It measures how similar two images appear to the human visual system.

4.2 Implementation Details

We augment the data during training by applying random horizontal flips and 90/180/270-degree rotations. For a fair comparison with the existing works, LR images are obtained through bi-cubic down-sampling from HR images. Empirically, the number of SCATBs in ML-CrAIST is set to 5. Also, the attention head number is set to 4, and the window size is set to 8. We train the model using the Adam optimizer with a batch size of 32 for 1000K iterations, starting with an initial learning rate of 10^{-4}, which is decreased by half after every 200k iterations. During each training iteration, LR patches of size 64×64 are randomly cropped as input. We have set the number of channels 64 in each convolution layer for ML-CrAIST (Ours). The proposed work is implemented using PyTorch, and all experimentations are performed on a single NVIDIA V100 GPU. Figure 5(b) shows the convergence of the model that we observed.

In our lighter version of ML-CrAIST (Ours-Li), we have used the same architecture shown in Figure 1 with a reduced number of channels in each convolution layer from 64 to 48.

Table 1. PSNR and SSIM comparison with the state-of-the-art on five datasets. Best, second best , and third best performance are presented in **red**, blue, and *green*.

			#params	FLOPs	**Set5**		**Set14**		**B100**		**Urban100**		**Manga109**	
Method		Years	(K)	(G)	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
VDSR	2×	CVPR'16	666	613	36.66	0.9542	33.05	0.9127	31.90	0.8960	30.76	0.9140	37.22	0.9750
MemNet		ICCV'17	678	2662.4	37.78	0.9597	33.28	0.9142	32.08	0.8978	31.31	0.9195	37.72	0.9740
SRMDNF		CVPR'18	1511	-	37.79	0.960	33.32	0.915	32.05	0.8985	31.33	0.9204	38.07	0.9761
CARN		ECCV'18	1592	222.8	37.76	0.9590	33.52	0.9166	32.09	0.8978	31.92	0.9256	38.36	0.9765
IMDN		MM'19	694	158.8	38.00	0.9605	33.63	0.9177	32.19	0.8996	32.17	0.9283	38.88	0.9774
LatticeNet		ECCV'20	756	169.5	38.06	0.9610	33.78	0.9193	32.25	0.9005	32.43	0.9302	38.94	0.9774
SwinIR		ICCVW'21	878	195.6	38.14	0.9611	33.86	0.9206	*32.31*	*0.9012*	32.76	0.9340	39.12	0.9783
ESRT		CVPRW'22	677	191.4	38.03	0.9600	33.75	0.9184	32.25	0.9001	32.58	0.9318	39.12	0.9774
NGSwin		CVPR'23	998	140.4	38.05	0.9610	*33.79*	0.9199	32.27	0.9008	32.53	0.9324	38.97	0.9777
OmniSR		CVPR'23	772	147.2	**38.22**	*0.9613*	**33.98**	*0.9210*	**32.36**	0.9020	**33.05**	0.9363	**39.28**	*0.9784*
Ours-Li			**743**	**97.2**	*38.15*	0.9615	33.64	0.9213	32.35	0.9020	*32.93*	*0.9361*	*39.23*	0.9785
Ours			1259	165.7	38.19	**0.9617**	33.77	**0.9220**	**32.36**	**0.9022**	33.04	**0.9370**	39.26	**0.9786**
VDSR	3×	CVPR'16	666	613	33.66	0.9213	29.77	0.8314	28.82	0.7976	27.14	0.8279	32.01	0.9340
MemNet		ICCV'17	678	2662.4	34.09	0.9248	30.00	0.8350	28.96	0.8001	27.56	0.8376	32.51	0.9369
EDSR		CVPRW'17	1555	160.2	34.37	0.9270	30.28	0.8417	29.09	0.8052	28.15	0.8527	33.45	0.9439
SRMDNF		CVPR'18	1528	-	34.12	0.9254	30.04	0.8382	28.97	0.8025	27.57	0.8398	33.00	0.9403
CARN		ECCV'18	1592	118.8	34.29	0.9255	30.29	0.8407	29.06	0.8034	28.06	0.8493	33.50	0.9440
IMDN		MM'19	703	56.3	34.36	0.9270	30.32	0.8417	29.09	0.8046	28.17	0.8519	33.61	0.9445
RFDN-L		ECCV'20	633	65.6	34.47	0.9280	30.35	0.8421	29.11	0.8053	28.32	0.8547	33.78	0.9458
LatticeNet		ECCV'20	765	76.3	34.40	0.9272	30.32	0.8416	29.10	0.8049	28.19	0.8513	33.63	0.9442
SwinIR		ICCVW'21	886	87.2	34.62	*0.9289*	30.54	0.8463	*29.20*	0.8082	28.66	0.8624	33.98	0.9478
ESRT		CVPRW'22	770	96.4	34.42	0.9268	30.43	0.8433	29.15	0.8063	28.46	0.8574	33.95	0.9455
NGSwin		CVPR'23	1007	66.6	34.52	0.9282	*30.53*	0.8456	29.19	0.8078	28.52	0.8603	33.89	0.9470
OmniSR		CVPR'23	780	74.4	**34.70**	0.9294	**30.57**	*0.8469*	29.28	*0.8094*	28.84	0.8656	*34.22*	*0.9487*
Ours-Li			**749**	**49.6**	*34.58*	0.9294	30.23	0.8474	29.28	0.8106	*28.73*	*0.8651*	34.26	0.9492
Ours			1268	84.1	**34.70**	**0.9302**	30.39	**0.8488**	**29.31**	**0.8111**	**28.89**	**0.8676**	**34.42**	**0.9501**
VDSR	4×	CVPR'16	666	613	31.35	0.8838	28.01	0.7674	27.29	0.7251	25.18	0.7524	28.83	0.8870
MemNet		ICCV'17	678	2662.4	31.74	0.8893	28.26	0.7723	27.40	0.7281	25.50	0.7630	29.42	0.8942
EDSR		CVPRW'17	1518	114.0	32.09	0.8938	28.58	0.7813	27.57	0.7357	26.04	0.7849	30.35	0.9067
SRMDNF		CVPR'18	1552	-	31.96	0.8925	28.35	0.7787	27.49	0.7337	25.68	0.7731	30.09	0.9024
CARN		ECCV'18	1592	90.9	32.13	0.8937	28.60	0.7806	27.58	0.7349	26.07	0.7837	30.47	0.9084
IMDN		MM'19	715	40.9	32.21	0.8948	28.58	0.7811	27.56	0.7353	26.04	0.7838	30.45	0.9075
RFDN-L		ECCV'20	643	37.4	32.28	0.8957	28.61	0.7818	27.58	0.7363	26.20	0.7883	30.61	0.9096
LatticeNet		ECCV'20	777	43.6	32.30	0.8962	28.68	0.7830	27.62	0.7367	26.25	0.7873	30.54	0.9075
SwinIR		ICCVW'21	897	49.6	32.44	*0.8976*	28.77	0.7858	27.69	0.7406	26.47	0.7980	30.92	*0.9151*
ESRT		CVPRW'22	751	67.7	32.19	0.8947	*28.69*	0.7833	27.69	0.7379	26.39	0.7962	30.75	0.9100
NGSwin		CVPR'23	1019	36.4	32.33	0.8963	**28.78**	*0.7859*	27.66	0.7396	26.45	0.7963	30.80	0.9128
OmniSR		CVPR'23	792	37.8	**32.49**	**0.8988**	**28.78**	*0.7859*	*27.71*	*0.7415*	26.64	*0.8018*	*31.02*	*0.9151*
Ours-Li			**758**	**25.5**	32.15	0.8962	28.40	0.7863	27.73	0.7426	*26.53*	0.8019	31.11	0.9162
Ours			1280	42.9	*32.36*	0.8984	28.53	**0.7895**	**27.78**	**0.7446**	**26.68**	**0.8057**	**31.17**	**0.9176**

4.3 Comparisons with the SOTA

To validate the superiority of ML-CrAIST, we compare it against recent state-of-the-art methods (SOATs) under a scale factor of 2, 3, and 4, respectively. In particular, former works, VDSR [15], MemNet [35], EDSR [21], SRMDNF [44], CARN [1], IMDN [14], RFDN-L [23], LatticeNet [25], SwinIR [20], ESRT [24], NGSwin [7], and OmniSR [37] are introduced for comparison.

■ **Quantitative results.** The quantitative results are presented in Table 1. In order to be fair comparison throughout the evaluation process, all models undergo training and testing processes using the same dataset. It is clear from the results that our method achieves the highest performance across all testing datasets. Compared to [37], ML-CrAIST has 0.20 dB improvement on Manga109 (×3). Also, we noticed that our method demonstrates the most significant improvement on B100, Urban100, and Manga109 datasets compared to existing methods, indicating its suitability for images rich in textured regions, geometric structures, and finer details of SR images. As shown in Table 2, we obtain a lower LPIPS score, suggesting a higher perceptual quality of the SR image. It is worth noting that by incorporating the frequency details and analyzing the features in multiple scales, ML-CrAIST surpasses the performance of the existing methods. Additionally, in Table 1, we have shown the results of our lighter method (**Ours-Li**) with reduced parameters and FLOPs. It takes the minimum FLOPs among all the existing schemes with comparable results. The FLOPs are ∼ 1.5× lesser than NGSwin with 1.01% and 0.37% PSNR and SSIM gain on Manga109 (4×). Further, We have shown the computational overhead during inference in Table 3.

Table 2. LPIPS score Comparison on 4×. Best performance is presented in **red**. Lower score is better.

Model	**Set5**	**Set14**	**B100**	**Urban100**	**Manga109**
IMDN	0.1317 ± 0.0659	0.1242 ± 0.0866	0.1907 ± 0.0601	0.0131 ± 0.0124	0.0038 ± 0.0032
SwinIR	**0.1287 ± 0.0642**	0.1209 ± 0.0870	0.1857 ± 0.0596	0.0111 ± 0.0106	0.0033 ± 0.0026
NGSwin	0.1291 ± 0.0640	0.1210 ± 0.0869	0.1861 ± 0.0595	0.0109 ± 0.0101	0.0035 ± 0.0029
OmniSR	0.1293 ± 0.0641	0.1193 ± 0.0848	0.1829 ± 0.0595	0.0102 ± 0.0093	0.0034 ± 0.0029
Ours-Li	0.1354 ± 0.0651	0.1197 ± 0.0859	0.1842 ± 0.0595	0.0105 ± 0.0097	0.0033 ± 0.0028
Ours	0.1312 ± 0.0642	**0.1173 ± 0.0845**	**0.1812 ± 0.0591**	**0.0101 ± 0.0094**	**0.0032 ± 0.0027**

Table 3. Single image inference time for 2×, 3×, and 4×, respectively

			Inference time (second)		
scale	Input dimension	Output dimension	OmniSR	Ours-Li	Ours
2×	(512, 382)	(1024, 764)	4.98	**4.24**	5.99
3×	(341, 254)	(1023, 762)	2.73	**2.12**	3.35
4×	(256, 191)	(1024, 764)	2.05	**1.94**	2.61

- **Visual Comparison.** Figure 2 shows the visual comparison of our method with SOTAs at ×2, ×3, and ×4 scales. It is observable that the HR images generated by ML-CrAIST exhibit more fine-grained details, whereas other methods produce blurred edges or artifacts in complex regions. For example, in the third image of Figure 2, our model can pleasantly restore the precise texture of the road. The visual results demonstrate that incorporating frequency information and analyzing features across multiple scales enables us to capture more structural information, preserve the geometric structure of the image, and generate realistic HR results.

4.4 Ablation Study

In this subsection, we perform a set of experimentations to exhibit the efficacy of ML-CrAIST in different settings.

- **Number of SCATBs.** Experimentally, we have set the number of SCATBs to 5. We also analyze the model performance by varying the SCATB number N. As depicted in Figure 4, compared to the smallest number of SCATB, increasing the number of SCATB leads to performance gains. It can be seen

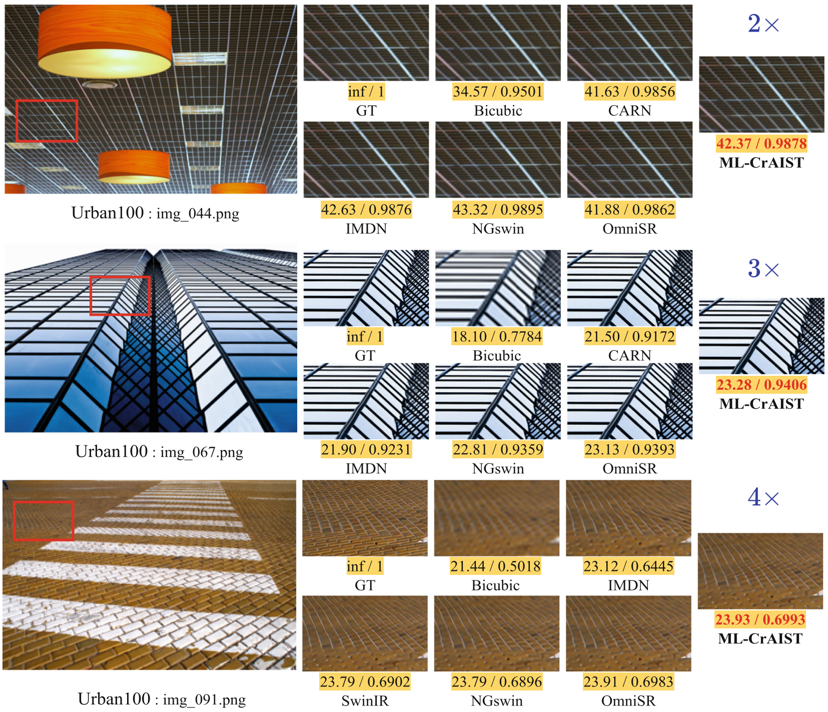


Fig. 2. Visual Comparison of our ML-CrAIST with the SOTA.

that ML-CrAIST with $N = 6$ or 7 produces a similar kind of result as $N = 5$ with higher parameters (refer to 4(d)).

■ **Effect of LHFIB.** We remove the frequency information and only take the spatial information to train our model. Figure 3(a) and 3(b) represent the diagram with and without the frequency information, respectively. The results are reported in the 5th row of the Table 4. The results of ML-CrAIST are superior with the frequency information, displaying that the frequency details can offer global dependency to enhance the representation capability of the model.

Table 4. Ablation studies with different settings of our model on 4×. Best result is represented in **red**.

	FLOPs	Set5		Set14		B100		Urban100		Manga109	
Model	(G)	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
w/o AFB (Addition)	42.80	32.28	0.8974	28.47	0.7886	27.56	0.7431	26.64	0.8041	31.14	0.9174
w/o AFB (Concatenation)	42.82	32.29	0.8974	28.45	0.7885	27.68	0.7435	26.63	0.8056	31.09	0.9174
DWT Level-1	41.11	32.15	0.8957	28.46	0.7872	27.72	0.7423	26.58	0.8022	31.04	0.9157
w/o CAB	41.79	32.31	0.8977	28.52	0.7881	27.76	0.7433	26.65	0.8043	31.10	0.9175
w/o LHFIB	42.53	32.29	0.8975	28.42	0.7888	27.26	0.7434	26.66	0.8050	31.11	0.9164
Full Model	42.91	**32.36**	**0.8984**	**28.53**	**0.7895**	**27.78**	**0.7446**	**26.68**	**0.8057**	**31.17**	**0.9176**

■ **Effect of CAB.** We execute experiments to investigate the significance of the CAB. Specifically, we compare the results of the model with and without CAB in the 4th row of Table 4. While removing the CAB, we used a simple element-wise addition operation. From the aspects of quantitative metrics, the use of CAB can obviously improve the SSIM and PSNR performance of the model. The visual comparison is shown in Figure 5(a).

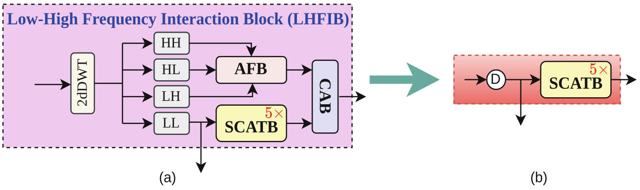


Fig. 3. (a) indicates the LHFIB, (b) indicates the diagram without frequency information. Ⓓ indicates the bi-cubic down-sampling operation.

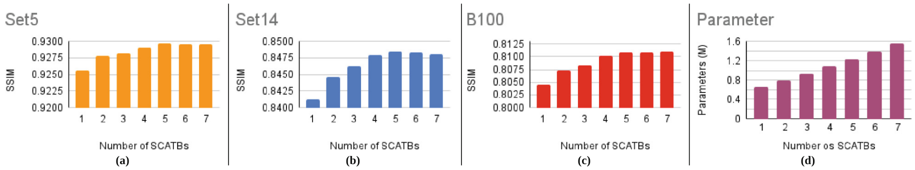


Fig. 4. (a), (b), and (c) refer the SSIM comparison, and (d) refers the number of parameters on 3× with different number of SCATBs.

- **Effect of AFB.** We explore the feature aggregation process in the 1^{st} and 2^{nd} row of Table 4. The results demonstrate that the proposed AFB produces promising outcomes compared to summation and concatenation methods.
- **Effect of multi-scale or multi-level DWT.** Third row of Table 4 justifies the importance of the 2-level 2dDWT or multi-scale analysis in our model.

Further, to validate each component of ML-CrAIST, in Figure 6, we have shown results in three different measurements: LPIPS, Blind/Referenceless Image Spatial Quality Evaluator (BRISQUE), and Edge Preservation Index (EPI). It can be seen that the full model has a lower LPIPS and BRISQUE and a high EPI value, which indicates that the image has fewer distortions, artifacts, and better edge preservation, aligns more closely with natural scene statistics, and is visually pleasing to human observers.

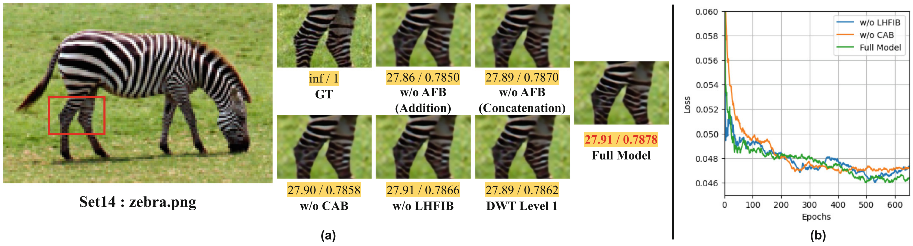


Fig. 5. (a) Visual comparison of different settings of ML-CrAIST. (b) Convergence graph of ML-CrAIST.

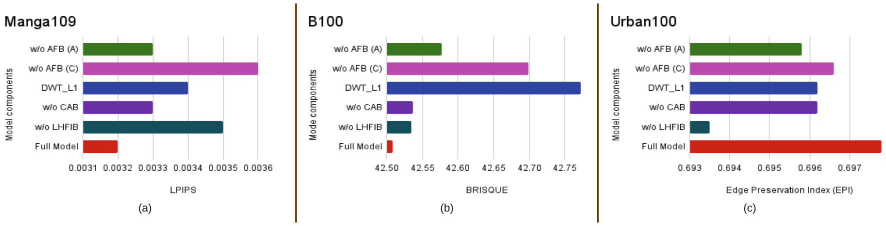


Fig. 6. LPIPS (↓), BRISQUE (↓), and EPI comparison between different components of ML-CrAIST. ↓ indicates lower is better.

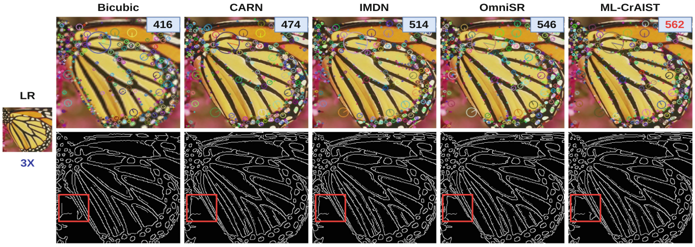


Fig. 7. Key-point and canny edge detection comparison between existing methods and ML-CrAIST. The top corner of the first row indicates the number of key points.

4.5 Impact on various application

To validate the practical applicability of our model, we employ ML-CrAIST as a prepossessing technique for image key-point detection and edge detection tasks, as shown in Figure 7. Initially, we employ scale-invariant feature transform (SIFT) to compute the key points. It can be observed that the key-point detection significantly increases after super-resolving the images using our method. Subsequently, we employ Canny edge detection to identify edges in the super-resolved images. Compared to the super-resolved image by SOTA models, our super-resolved image exhibits more localized edge features. In the second row of Figure 7, we have marked using a red box where our method captures edges perfectly, but others fail.

5 Conclusion

In this paper, we propose a transformer-based multi-scale super-resolution architecture called ML-CrAIST, demonstrating the advantage of modeling both spatial and frequency details for the SR task. Our cross-attention block seamlessly performs message passing between low and high-frequency features across multiple scales in the network and acknowledges their correlation. Furthermore, we propose AFB to effectively fuse the high frequency cubes, which boosts the overall performance. Finally, we validate the rationale and efficiency of the ML-CrAIST by conducting extensive experimentation across various benchmark datasets. We additionally conduct an ablation study to assess the impact of various configurations within ML-CrAIST.

References

1. Ahn, N., Kang, B., Sohn, K.A.: Fast, accurate, and lightweight super-resolution with cascading residual network. In: Proceedings of the European conference on computer vision (ECCV). pp. 252–268 (2018)
2. Bevilacqua, M., Roumy, A., Guillemot, C., Alberi-Morel, M.L.: Low-complexity single-image super-resolution based on nonnegative neighbor embedding. BMVC (2012)
3. Cao, H., Wang, Y., Chen, J., Jiang, D., Zhang, X., Tian, Q., Wang, M.: Swin-unet: Unet-like pure transformer for medical image segmentation. In: European conference on computer vision. pp. 205–218. Springer (2022)
4. Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., Zagoruyko, S.: End-to-end object detection with transformers. In: European conference on computer vision. pp. 213–229. Springer (2020)
5. Chen, H., Wang, Y., Guo, T., Xu, C., Deng, Y., Liu, Z., Ma, S., Xu, C., Xu, C., Gao, W.: Pre-trained image processing transformer. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 12299–12310 (2021)
6. Chen, X., Wang, X., Zhou, J., Qiao, Y., Dong, C.: Activating more pixels in image super-resolution transformer. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 22367–22377 (2023)

7. Choi, H., Lee, J., Yang, J.: N-gram in swin transformers for efficient lightweight image super-resolution. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2071–2081 (2023)
8. Chu, X., Zhang, B., Ma, H., Xu, R., Li, Q.: Fast, accurate and lightweight super-resolution with neural architecture search. In: 2020 25th International conference on pattern recognition (ICPR). pp. 59–64. IEEE (2021)
9. Dai, T., Cai, J., Zhang, Y., Xia, S.T., Zhang, L.: Second-order attention network for single image super-resolution. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 11065–11074 (2019)
10. Dong, C., Loy, C.C., He, K., Tang, X.: Image super-resolution using deep convolutional networks. IEEE Trans. Pattern Anal. Mach. Intell. **38**(2), 295–307 (2015)
11. Gao, Q., Zhao, Y., Li, G., Tong, T.: Image super-resolution using knowledge distillation. In: Asian Conference on Computer Vision. pp. 527–541. Springer (2018)
12. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. Advances in neural information processing systems **27** (2014)
13. Huang, J.B., Singh, A., Ahuja, N.: Single image super-resolution from transformed self-exemplars. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 5197–5206 (2015)
14. Hui, Z., Gao, X., Yang, Y., Wang, X.: Lightweight image super-resolution with information multi-distillation network. In: Proceedings of the 27th acm international conference on multimedia. pp. 2024–2032 (2019)
15. Kim, J., Lee, J.K., Lee, K.M.: Accurate image super-resolution using very deep convolutional networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1646–1654 (2016)
16. Kim, J., Lee, J.K., Lee, K.M.: Deeply-recursive convolutional network for image super-resolution. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1637–1645 (2016)
17. Kong, F., Li, M., Liu, S., Liu, D., He, J., Bai, Y., Chen, F., Fu, L.: Residual local feature network for efficient super-resolution. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 766–776 (2022)
18. Ledig, C., Theis, L., Huszár, F., Caballero, J., Cunningham, A., Acosta, A., Aitken, A., Tejani, A., Totz, J., Wang, Z., et al.: Photo-realistic single image super-resolution using a generative adversarial network. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4681–4690 (2017)
19. Li, X., Wang, W., Hu, X., Yang, J.: Selective kernel networks. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 510–519 (2019)
20. Liang, J., Cao, J., Sun, G., Zhang, K., Van Gool, L., Timofte, R.: Swinir: Image restoration using swin transformer. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 1833–1844 (2021)
21. Lim, B., Son, S., Kim, H., Nah, S., Mu Lee, K.: Enhanced deep residual networks for single image super-resolution. In: Proceedings of the IEEE conference on computer vision and pattern recognition workshops. pp. 136–144 (2017)
22. Liu, C., Yang, H., Fu, J., Qian, X.: Learning trajectory-aware transformer for video super-resolution. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5687–5696 (2022)
23. Liu, J., Tang, J., Wu, G.: Residual feature distillation network for lightweight image super-resolution. In: Computer Vision–ECCV 2020 Workshops: Glasgow, UK, August 23–28, 2020, Proceedings, Part III 16. pp. 41–55. Springer (2020)

24. Lu, Z., Li, J., Liu, H., Huang, C., Zhang, L., Zeng, T.: Transformer for single image super-resolution. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 457–466 (2022)
25. Luo, X., Xie, Y., Zhang, Y., Qu, Y., Li, C., Fu, Y.: Latticenet: Towards lightweight image super-resolution with lattice block. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXII 16. pp. 272–289. Springer (2020)
26. Martin, D., Fowlkes, C., Tal, D., Malik, J.: A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. In: Proceedings Eighth IEEE International Conference on Computer Vision. ICCV 2001. vol. 2, pp. 416–423. IEEE (2001)
27. Matsui, Y., Ito, K., Aramaki, Y., Fujimoto, A., Ogawa, T., Yamasaki, T., Aizawa, K.: Sketch-based manga retrieval using manga109 dataset. Multimedia Tools and Applications **76**, 21811–21838 (2017)
28. Mei, Y., Fan, Y., Zhou, Y.: Image super-resolution with non-local sparse attention. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3517–3526 (2021)
29. Niu, B., Wen, W., Ren, W., Zhang, X., Yang, L., Wang, S., Zhang, K., Cao, X., Shen, H.: Single image super-resolution via a holistic attention network. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XII 16. pp. 191–207. Springer (2020)
30. Park, S.H., Moon, Y.S., Cho, N.I.: Flexible style image super-resolution using conditional objective. IEEE Access **10**, 9774–9792 (2022)
31. Park, S.H., Moon, Y.S., Cho, N.I.: Perception-oriented single image super-resolution using optimal objective estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1725–1735 (2023)
32. Pramanick, A., Megha, D., Sur, A.: Attention-based spatial-frequency information network for underwater single image super-resolution. In: ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 3560–3564. IEEE (2024)
33. Rakotonirina, N.C., Rasoanaivo, A.: Esrgan+: Further improving enhanced super-resolution generative adversarial network. In: ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 3637–3641. IEEE (2020)
34. Tai, Y., Yang, J., Liu, X.: Image super-resolution via deep recursive residual network. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3147–3155 (2017)
35. Tai, Y., Yang, J., Liu, X., Xu, C.: Memnet: A persistent memory network for image restoration. In: Proceedings of the IEEE international conference on computer vision. pp. 4539–4547 (2017)
36. Timofte, R., Agustsson, E., Van Gool, L., Yang, M.H., Zhang, L.: Ntire 2017 challenge on single image super-resolution: Methods and results. In: Proceedings of the IEEE conference on computer vision and pattern recognition workshops. pp. 114–125 (2017)
37. Wang, H., Chen, X., Ni, B., Liu, Y., Liu, J.: Omni aggregation networks for lightweight image super-resolution. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 22378–22387 (2023)
38. Wang, X., Yu, K., Wu, S., Gu, J., Liu, Y., Dong, C., Qiao, Y., Change Loy, C.: Esrgan: Enhanced super-resolution generative adversarial networks. In: Proceedings of the European conference on computer vision (ECCV) workshops. pp. 0–0 (2018)

39. Wang, Z., Cun, X., Bao, J., Zhou, W., Liu, J., Li, H.: Uformer: A general u-shaped transformer for image restoration. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 17683–17693 (2022)
40. Wu, B., Xu, C., Dai, X., Wan, A., Zhang, P., Yan, Z., Tomizuka, M., Gonzalez, J., Keutzer, K., Vajda, P.: Visual transformers: Token-based image representation and processing for computer vision. arXiv preprint arXiv:2006.03677 (2020)
41. Zamir, S.W., Arora, A., Khan, S., Hayat, M., Khan, F.S., Yang, M.H.: Restormer: Efficient transformer for high-resolution image restoration. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 5728–5739 (2022)
42. Zeyde, R., Elad, M., Protter, M.: On single image scale-up using sparse-representations. In: Curves and Surfaces: 7th International Conference, Avignon, France, June 24-30, 2010, Revised Selected Papers 7. pp. 711–730. Springer (2012)
43. Zhang, D., Huang, F., Liu, S., Wang, X., Jin, Z.: Swinfir: Revisiting the swinir with fast fourier convolution and improved training for image super-resolution. arXiv preprint arXiv:2208.11247 (2022)
44. Zhang, K., Zuo, W., Zhang, L.: Learning a single convolutional super-resolution network for multiple degradations. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3262–3271 (2018)
45. Zhang, X., Zeng, H., Guo, S., Zhang, L.: Efficient long-range attention network for image super-resolution. In: European Conference on Computer Vision. pp. 649–667. Springer (2022)
46. Zhang, Y., Chen, H., Chen, X., Deng, Y., Xu, C., Wang, Y.: Data-free knowledge distillation for image super-resolution. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7852–7861 (2021)
47. Zhang, Y., Li, K., Li, K., Wang, L., Zhong, B., Fu, Y.: Image super-resolution using very deep residual channel attention networks. In: Proceedings of the European conference on computer vision (ECCV). pp. 286–301 (2018)

Attentive Color Fusion Transformer Network (ACFTNet) for Underwater Image Enhancement

Mohd Ubaid Wani[1](✉), Md Raqib Khan[2], Ashutosh Kulkarni[1],
Shruti S. Phutke[3], Santosh Kumar Vipparthi[1], and Subrahmanyam Murala[2]

[1] CVPR Lab, Indian Institute of Technology Ropar, Bara Phool, India
2022eem1008@iitrpr.ac.in
[2] Trinity College Dublin, Dublin, Ireland
[3] Emerging Technologies and Innovation Lab,
Yamaha Motor Solutions, Faridabad, India

Abstract. Underwater imagery often suffers from issues like color distortion, haze, and reduced visibility due to light's interaction with water, posing challenges for applications like autonomous underwater vehicles. To address these obstacles effectively, we introduce the Attentive Color Fusion Transformer Network (ACFTNet) for underwater image enhancement. At the core of our proposal lies a novel Adaptive Dual-Gated Attentive Fusion Block (ADGAFB), which seamlessly integrates localized transmission features and global illumination characteristics. Subsequently, it employs a dual-gated mechanism to generate attentive features for each channel (R, G, and B). To ensure accurate color fidelity, we introduce the Color-Attentive Fusion Block. This block adeptly merges attentive features obtained from each R, G, and B channel, ensuring precise color representation. To selectively transmit features from the encoder to the corresponding decoder, we utilize an Adaptive Kernel-Based Channel Attention Module. Moreover, within the transformer block, we propose a Multi-Receptive Field Feed-Forward Gated Network to further refine the restoration process. Through comprehensive evaluations on benchmark synthetic (UIEB, EUVP) and real-world (UIEB (challenging-60), UCCS, U45) underwater image datasets, our method exhibits superior performance, as verified by extensive ablation studies and comparative analyses. The testing code is available at https://github.com/MohdUbaidwani/ACFTNet.

Keywords: Underwater Image Enhancement · Adaptive Dual-Gated Attentive Fusion Block · Transformer

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/978-3-031-78305-0_20.

A. Antonacopoulos et al. (Eds.): ICPR 2024, LNCS 15321, pp. 308–324, 2025.
https://doi.org/10.1007/978-3-031-78305-0_20

1 Introduction

Enhancing underwater imaging is crucial for effective observation of marine environments, aiding in the identification and visibility of objects and marine life. Such algorithms find applications in diverse fields like underwater mine detection, inspection, surveillance, autonomous vehicles, and robotics systems [11,51]. Despite advancements in imaging technology, underwater environments present significant challenges, including poor equipment quality [27], illumination, scattering, and absorption [41], resulting in degraded images with color casts, low visibility, and contrast issues. These obstacles hinder the accuracy of underwater applications, necessitating advanced algorithms tailored to enhance image quality for more effective exploration and application across diverse fields.

Underwater image enhancement (UIE) methods typically fall into three categories. The first approach involves physically modeling the underwater environment, like computing transmission maps [5,10], but they are limited by underwater complexities. In the second category, visual prior-based methods [1,23] focus on adjusting pixel values for contrast and brightness to enhance perceptual quality, yet they overlook the physical deterioration process.

Deep learning methods [8,9,24,28] constitute the third category, showing significant progress in underwater image enhancement by addressing UIE challenges. Transformers [48] have further enhanced performance metrics by leveraging global dependencies in vision tasks. Underwater imaging relies on spatial features like pixel values and positions, and spectral features such as fine details and consistent patterns, where features from R, G, and B channels are treated differently based on their wavelengths. Processing both spatial and spectral features is crucial for improving visibility, color effectiveness, and handling noisy spikes caused by light scattering from suspended particles in water. Recognizing the pivotal role of underwater image enhancement as the foundational step in underwater vision tasks, our goal is to enhance the visibility of underwater images. With this objective, we introduce the Attentive Color Fusion Transformer Network (ACFTNet) for underwater image enhancement. Unlike Spectroformer [20], the proposed ACFTNet harnesses color-wise spectral and spatial features to improve color fidelity and capture essential details in underwater images. Our proposed method, ACFTNet, addresses various degradation factors in underwater images such as scattering, color casts, and absorption-based transmission maps. It incorporates essential information from separate R, G, and B color spaces at both local and global levels, encompassing both spatial and frequency domains, thereby mitigating various underwater degradation factors. We propose the Adaptive Dual-Gated Attentive Fusion Block (ADGAFB) seen in Fig. 1, seamlessly integrating localized transmission features and global illumination characteristics. Subsequently, it employs a dual-gated mechanism to generate attentive features for each channel (R, G, and B), leading to efficient learning of fine details and the structural details of degraded underwater images. To uphold color fidelity, we introduce a novel Color Attentive Fusion Block. This block maintains color fidelity by integrating color-wise attentive features from the R, G, and B channel features obtained from the Adaptive Dual-Gated Attentive Fusion Block (ADGAFB) with the original RGB features at each level. This approach effectively addresses challenges related to color distortion and contrast reduction. To enhance the recovery process, encoder features are typically relayed to decoder features via direct skip connections [54]. However, this direct forwarding can degrade the decoder's performance, leading to inefficiencies in image enhancement [19]. To address this, we employ an Adaptive Kernel-Based Channel Attention Module, selectively enhancing the transmission of encoder features to decoder features, as shown in Fig. 1. Furthermore, within the transformer block, we introduce a Multi-Receptive Field Feed-Forward Gated Network to transmit the most important multi-receptive and diverse information from the input-degraded side to the output-enhancing side, facilitating better refinement and reconstruction of textures and periodic patterns. The main contributions of this work can be summarized as:

- A novel Attentive Color Fusion Transformer Network (ACFTNet) architecture is proposed for underwater image enhancement.
- We propose the Adaptive Dual-Gated Attentive Fusion Block (ADGAFB), seamlessly integrating localized transmission features and global illumination characteristics. Subsequently, it employs a dual-gated mechanism to generate attentive features for each channel (R, G, and B), leading to efficient learning of fine details and the structural details of degraded underwater images.
- We propose Color Attentive Fusion Block, which attentively combines the output feature of the encoder's transformer block with the output of (ADGAFB) for each separate R, G, B feature, resulting in comprehensive enhancement of degraded underwater images through efficient color feature acquisition.
- Moreover, within the transformer block, we propose a Multi-Receptive Field Feed-Forward Gated Network facilitating refined texture and periodic pattern reconstruction by transmitting crucial multi-receptive and diverse information from the degraded input to the enhancing output side.

The ablation study examines various configurations of the proposed approach. We conduct multiple experiments to demonstrate the effectiveness of the proposed method on both synthetic and real-world underwater images.

2 Related Work

2.1 Underwater Image Enhancement

Enhancing underwater images is vital for advanced computer vision tasks like object detection, recognition, and tracking. Existing methods for Underwater Image Enhancement (UIE) are categorized into four main groups: hardware-dependent, physical model-dependent, non-physical model-dependent, and deep learning-dependent methods.

Hardware-based Methods: For underwater image enhancement, some approaches employed specialized hardware, stereo-vision techniques, and polarization filters [40,47]. However, these approaches come with certain drawbacks.

Hardware-based methods can be costly and complex, introducing challenges for widespread adoption. Methods that rely on multiple images may be unsuitable for real-time applications. In contrast, [6] single-image enhancement stands out for challenging underwater scenes, offering a distinctive approach to address the complexities of underwater imaging.

Physical or Optical Methods: These methods rely on underwater image formation models, where the degraded image quality depends on the transmission map and underwater backlighting. Yang et al. [52] introduced a modified algorithm using the dark channel prior, while Chiang et al. [4] enhanced it with wavelength-dependent compensation. The Underwater Dark Channel Prior (UDCP) [12] specifically addresses red channel unreliability. Additionally, Peng et al. [38] proposed the Generalized Dark Channel Prior (GDCP) for adaptive color correction and comprehensive image restoration.

Non-Physical model based Methods: These methods aim to spatially enhance degraded underwater images by adjusting pixel values. Iqbal et al. [14] widened the pixel range in RGB and HSV color spaces to amplify contrast and saturation in underwater images. Ancuti et al. [2] introduced a fusion technique blending contrast-enhanced and color-corrected images using a multi-scale approach. Fu et al. [34] proposed a retinex-based method incorporating color correction, layer decomposition, and overall enhancement.

Deep Learning approaches: With the continuous advancement of deep learning, computer vision tasks have seen significant performance improvements. Supervised methods like UWCNN [25], WaterNet [46], and Spectroformer [20] employ deep learning for underwater image enhancement. UWCNN uses deep-supervised learning, while WaterNet utilizes a gated fusion network with gamma-corrected, contrast-enhanced, and white-balanced inputs. Spectroformer incorporates a multi-query-based attention mechanism. For unsupervised methods, Jiang et al. [17] proposed a perceptual adversarial network with adaptive feature fusion, and Li et al. [28] introduced WaterGAN, generating underwater-style images without paired training samples. Yang et al. [53] introduced a conditional generative adversarial network (cGAN) to enhance underwater image visual quality significantly. Du, D., et al. [7] proposed UIE with Diffusion Prior (UIEDP), a novel framework treating UIE as a posterior distribution sampling process of clear images conditioned on degraded underwater inputs. Their probabilistic nature, similar to Markov processes, adapts to local image gradients, enhancing versatility in image enhancement tasks [45].

2.2 Transformers in Computer Vision

Transformers based attention mechanisms, revolutionize computer vision by efficiently capturing long-range dependencies in input features, making them ideal for various tasks from low to high-level vision. Specifically tailored for restoration tasks like image denoising, deraining, and deblurring, Restormer [54] presents an efficient transformer network. Peng et al. [37] propose a U-shaped transformer for enhancing underwater images. Kong et al. [22] introduce an alternative Transformer-based approach for high-quality image deblurring, streamlining attention using frequency-domain characteristics.

3 Proposed Work

Our primary objective is to improve the quality of underwater degraded images by employing techniques from both the spatial [19] and frequency domains [49]. Our approach aims to retain fine details such as edges, enhance color components, improve contrast, and eliminate undesirable noise artifacts caused by light scattering. Now, let's delve into the primary pipeline of our proposed network, ACFTNet depicted in Fig. 1. After that, we provide more details of the individual proposed blocks.

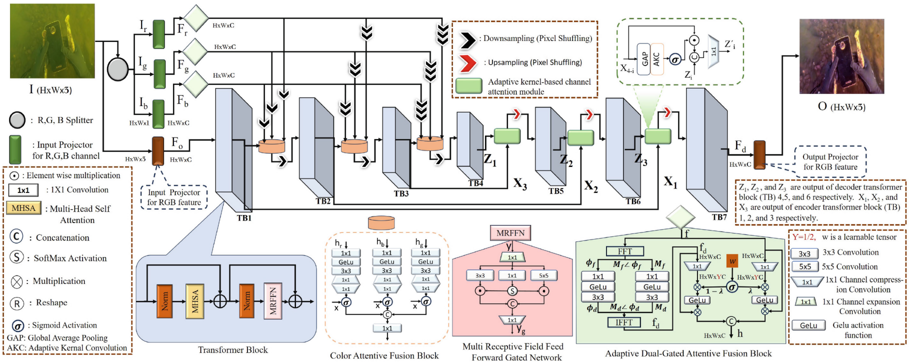

I (HxWx3)
: R,G, B Splitter
: Input Projector for R,G,B channel
I_r
I_g
I_b
F_r
F_g
F_b
F_o
HxWxC
HxWx1
HxWx3
Input Projector for RGB feature
⊙ : Element wise multiplication
1x1 : 1X1 Convolution
MHSA : Multi-Head Self Attention
Ⓒ : Concatenation
Ⓢ : SoftMax Activation
⊗ : Multiplication
Ⓡ : Reshape
σ : Sigmoid Activation
GAP: Global Average Pooling
AKC: Adaptive Kernal Convolution
TB1
TB2
TB3
TB4
TB5
TB6
TB7
: Downsampling (Pixel Shuffling)
: Upsampling (Pixel Shuffling)
Adaptive kernel-based channel attention module
X_{4-i}
GAP
AKC
Z_i
Z'_i
Z_1
Z_2
Z_3
X_1
X_2
X_3
F_d
O (HxWx3)
Output Projector for RGB feature
Z_1, Z_2, and Z_3 are output of decoder transformer block (TB) 4,5, and 6 respectively. X_1, X_2, and X_3 are output of encoder transformer block (TB) 1, 2, and 3 respectively.
Norm
MHSA
MRFFN
Transformer Block
h_r
h_b
h_g
GeLu
3x3
x
Color Attentive Fusion Block
MRFFN
y
y_g
5x5
Multi Receptive Field Feed Forward Gated Network
f
FFT
IFFT
ϕ_f
$M_f \angle \phi_f$
M_f
ϕ_d
$M_d \angle \phi_d$
M_d
f_d
w
HxWxYC
$1-\lambda$
λ
h
Adaptive Dual-Gated Attentive Fusion Block
Y=1/2, w is a learnable tensor
3x3 Convolution
5x5 Convolution
1x1 Channel compression Convolution
1x1 Channel expansion Convolution
Gelu activation function

Fig. 1. The architectural schematic of ACFTNet for underwater image enhancement includes four key components: **Adaptive Dual-Gated Attentive Fusion Block (ADGAFB), Color Attentive Fusion Block, Multi-Receptive Field Feed Forward Gated Network, and Adaptive kernel- based channel attention module.** ADGAFB combines spatial and frequency domain features to enhance underwater imaging, addressing challenges like scattering, absorption, and structural pattern loss, it operates on individual R, G, and B channels, producing color-dependent enhancements. The Color Attentive Fusion Block combines attentive features by R, G, and B. The Multi-Receptive Field Feed Forward Gated Network refines the enhancement process within the transformer block. Adaptive Kernel-based Channel Attention Module, enhances feature transmission between encoder and decoder, improving performance and feature augmentation.

Overall Pipeline The initial step involves decomposing the degraded input RGB image $\mathbf{I}$ into separate $\mathbf{I}_r$, $\mathbf{I}_g$, and $\mathbf{I}_b$ channels, as depicted in Fig. 1. Subsequently, convolution operations are applied to $\mathbf{I}_r$, $\mathbf{I}_g$, $\mathbf{I}_b$ channels and the degraded input RGB image $\mathbf{I}$, resulting in the generation of shallow features $\mathbf{F}_o$, $\mathbf{F}_r$, $\mathbf{F}_g$, $\mathbf{F}_b \in \mathbb{R}^{H\times W\times C}$ corresponding to $\mathbf{I}$, $\mathbf{I}_r$, $\mathbf{I}_g$, and $\mathbf{I}_b$, respectively. The shallow features $\mathbf{F}_r$, $\mathbf{F}_g$, and $\mathbf{F}_b$ are processed by Adaptive Dual-Gated Attentive Fusion Block (ADGAFB), smoothly and effectively integrating localized transmission features and global illumination characteristics. Subsequently, it employs a dual-gated mechanism to generate attentive features for each channel (R, G, and B). Meanwhile, $\mathbf{F}_o$ undergoes processing in the Transformer Block which consists of multi-head self-attention and novel Multi-Receptive Field Feed Forward Gated Network, to extract features as depicted in Fig. 1. Color Attentive Fusion Block attentively combines the output feature of the encoder's transformer block with the output of (ADGAFB) for each separate R, G, and B feature. An Adaptive Kernel-based Channel Attention Module is used to transmit attentive information from the encoder to the respective decoder, enhancing the effectiveness of the underwater image. We utilize pixel-unshuffle and pixel-shuffle operations [43] for downsampling and upsampling features, respectively. At the end, a convolution layer is applied to the resulting deep features $\mathbf{F}_d \in \mathbb{R}^{H\times W\times C}$ to produce the final output, resulting in the enhanced image $\mathbf{O}$.

3.1 Adaptive Dual-Gated Attentive Fusion Block

To address challenges in underwater image processing, such as loss of fine details and visibility due to scattering and absorption, we introduce the Adaptive Dual-Gated Attentive Fusion Block (ADGAFB). Traditional spatial domain-only methods often lack a holistic understanding of the global image structure [30]. Exploiting frequency domain features, which provide advantages like illumination-invariant representations and capturing global information [33], our approach incorporates a dual-gated attentive fusion mechanism.

Initially, to extract frequency domain features to capture essential global information, illumination invariance, and fine details associated with high-frequency components [44], we apply FFT (Fast Fourier Transform) on feature f and split it into its frequency domain components: ϕ_f and magnitude component M_f. These components then sequentially pass through a 1×1 convolution, GeLu activation function, and a 3×3 convolution block. The learned phase (ϕ_d) and magnitude (M_d) are then passed to an IFFT (Inverse Fast Fourier Transform) block, resulting in the feature transformed back into the spatial domain f_d. Subsequently, the Adaptive Dual-Gated Attentive Fusion Block (ADGAFB) is applied, integrating both spatial and frequency domain features f and f_d respectively, thereby enhancing the overall representation. The adaptive nature of the fusion block allows for dynamic adjustment of the importance of spatial and frequency domain features based on input image characteristics, ensuring robustness across diverse underwater conditions. By amalgamating the strengths of spatial and frequency domain features through the ADGAFB, our proposed method aims to elevate the quality of underwater image processing, mitigating challenges posed by scattering, absorption, and loss of structural patterns. ADGAFB operates on each R, G, and B channel feature separately, generating color-dependent enhancements for underwater images, as depicted in Fig. 1. Mathematically, ADGAFB can be expressed as:

$$\phi_f, M_f = FFT(f) \tag{1}$$

$$\phi_{f_d}, M_{f_d} = \psi_3(GeLu(\psi_1(\phi_f))), \psi_3(GeLu(\psi_1(M_f))) \tag{2}$$

$$f_d = IFFT(\phi_{f_d}, M_{f_d}) \tag{3}$$

$$f' = \langle GeLu(\lambda_1 * \psi_1(f_d)) * f_d, GeLu(\lambda_2 * \psi_1(f)) * f \rangle \tag{4}$$

where $< \cdot >$ represents the concatenation operation, $\lambda_2 + \lambda_1 = 1$, and $\lambda_i = \sigma(w)$, $i \in (1, 2)$. (ψ_m) denotes a convolutional layer with kernel size $m \times m$, f' represents the output of ADGAFB, and f and f_d are the original spatial feature and the frequency domain learned feature, respectively. λ represents the mixup weighting parameter [55].

3.2 Color-Attentive Fusion Block

This module plays a crucial role in enriching the color characteristics of the red (**R**), green (**G**), and blue (**B**) channels. In underwater environments, issues like scattering and under-illumination often result in diminished color visibility and color cast problems. To address these challenges, fine and color-riched details from the ADGAFB are fed into this block during subsequent stages of processing. Initially, the 1×1 convolutions extract global color information separately from the incoming **R**, **G**, and **B** features. These outputs then pass through the GeLu activation function, aiding the model in capturing non-uniformities within the color context. To preserve the color correlation of each channel with the original RGB stream, the enhanced color components are providing the attention to the RGB features using the sigmoid function, resulting attentive feature are fused depicted in Fig. 1. This mechanism ensures a natural color touch to the degraded images. Finally, for further fine-tuning of color characteristics, a 1×1 convolutional layer is applied. This process facilitates the refinement of color attributes, contributing to the overall enhancement of underwater image quality.

3.3 Multi-Receptive Field Feed Forward Gated Network

Underwater images contain a variety of entities, including fishes, coral reefs, underwater terrain, and man-made structures, each distinguished by their forms and scales. A prominent challenge in underwater image enhancement involves the processing of these complex and diverse entities [32,35,36]. Traditional methods often rely on kernels of fixed dimensions to account for variations in size and shape, which restricts their ability to extract multi-scale information. In light of this, we propose a Multi-Receptive Field Feed Forward Gated Network (see MRFFN in Fig. 1). The proposed MRFFN can be detailed mathematically as:

$$y_g(y) = \psi_3 \left(\left\langle \wp \left(\psi_3^d(\theta), \psi_1^d(\theta) \right), \wp \left(\psi_5^d(\theta), \psi_1^d(\theta) \right) \right\rangle \right) \tag{5}$$

$$\wp(a, b) = a * \zeta(b) \tag{6}$$

$$\theta = \psi_1 \left(y \right) \tag{7}$$

where, $\langle a, b \rangle$ represents the concatenation of a, b, $\wp(\cdot)$ is the gating operation, $\zeta(\cdot)$ is Softmax activation. Such multi-receptive learning captures the various objects in the underwater images with various shapes and sizes to ensure meaningful enhancement of the images. The effectiveness of the proposed MRFFN block is analyzed in the ablation study (Sec. 5).

3.4 Adaptive Kernel-based Channel Attention Module

In an encoder-decoder UNet-like structure aimed at facilitating the recovery process, encoder features are commonly relayed directly to corresponding decoder features via skip connections [54]. Nevertheless, this direct feature forwarding can inadvertently introduce degradation to the decoder, resulting in inefficiencies in generating a fully enhanced image [19]. To mitigate this challenge, we utilize an Adaptive Kernel-Based Channel Attention Module. This module selectively enhances the transmission of encoder features to corresponding decoder features, as illustrated in Fig. 1. Mathematically, it can be explained as:

$$Z_i = \phi_1(\langle Z_i, \sigma(\varphi(GAP(X_{N-i}))) * X_{N-i}\rangle) \tag{8}$$

where, $< \cdot >$ is concatenation operation, GAP Global Average Pooling, φ is Adaptive Kernal Convolution, $\sigma(\cdot)$ Sigmoid activation, $N = 4$ and $i \in (1, 2, 3)$. Attentively transferring the feature(X_{N-i}) from the encoder aids the network in sharing essential information for the reconstruction process. Additionally, ϕ_1 is employed to halve the channels of the combined attentive encoder feature and the respective decoder feature. This reduction facilitates the network in learning the effective constraints for feature consideration pertinent to the enhancement task.

3.5 Training Loss Functions

To train our proposed model, inspired by Spectroformer [20] we have integrated the following loss functions as illustrated in the equation below:

$$L_T = \alpha_1 L_C + \alpha_2 L_G + \alpha_3 L_M + \alpha_4 L_P \tag{9}$$

where $\alpha_1, \alpha_2, \alpha_3, \alpha_4 \in \{0.25, 0.3, 0.7, 0.9\}$ are the weight coefficients of various loss functions are set empirically. During training, we are using a total loss

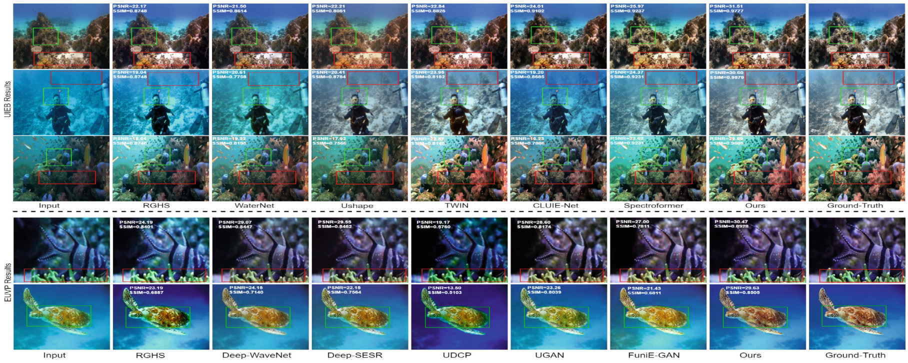


Fig. 2. ACFTNet achieves enhanced results on the UIEB and EUVP dataset: improved PSNR(dB) values, superior SSIM metrics, enhanced low-quality features with deep details, and efficient restoration of color casts.

Table 1. Comparative analysis of our proposed method (referred to as Ours) alongside established state-of-the-art techniques on the UIEB and EUVP datasets for enhancing underwater images. (↑ denotes higher is better, with **bold** representing **best**, and underline represents second best results).

Dataset	Method	Publications	PSNR ↑	SSIM ↑	UIQM ↑
UIEB	RGHS [13]	RAL-20	14.57	0.791	2.410
	WaterNet [26]	TIP-19	19.81	0.864	2.818
	CLUIE-Net [29]	TCSVT-22	20.37	0.890	2.674
	U-shape [37]	TIP-23	22.91	0.910	2.725
	TWIN [34]	TIP-22	23.72	0.830	3.024
	Spectroformer [20]	WACV-24	24.96	0.917	3.075
	Ours	- -	**26.02**	**0.931**	**4.792**
EUVP	UDCP[31]	CGA-16	20.31	0.519	3.702
	RGHS[13]	RAL-20	23.42	0.641	3.282
	FUniE-GAN[21]	JOET-22	26.22	0.790	**4.770**
	UGAN[9]	JOET-20	26.55	0.800	4.382
	Deep-SESR[15]	RSS-20	26.55	0.800	4.527
	Deep-WaveNet[42]	ACM-23	28.62	0.830	3.040
	ours	- -	**29.44**	**0.847**	4.298

Table 2. Comparative analysis of our proposed method and established state-of-the-art techniques on real-world datasets for underwater image restoration. (↑ denotes higher effectiveness, while ↓ indicates lower effectiveness).

Dataset	Method	UIQM ↑	UISM ↑	NIQE ↓	BRISQUE ↓
U45	RGHS[13]	2.506	5.558	3.8727	18.5190
	WaterNet[26]	3.091	6.187	4.5966	21.1563
	CLUIE-Net[29]	2.890	5.988	3.8743	20.6126
	U-shape[37]	2.923	5.567	4.3098	21.5656
	TWIN[34]	3.135	6.698	3.9929	20.0891
	Spectroformer[20]	3.243	**7.354**	3.8420	19.9573
	Ours	**4.405**	7.247	**3.780**	**18.104**
UCCS	RGHS[13]	2.506	5.558	4.209	26.360
	WaterNet[26]	3.134	6.187	6.104	24.275
	CLUIE-Net[29]	3.066	6.715	4.420	29.524
	U-shape[37]	2.874	5.391	4.401	23.549
	TWIN[34]	3.119	6.732	4.370	25.755
	Spectroformer[20]	3.209	6.563	**3.982**	**23.258**
	Ours	**4.375**	**6.953**	4.264	27.042

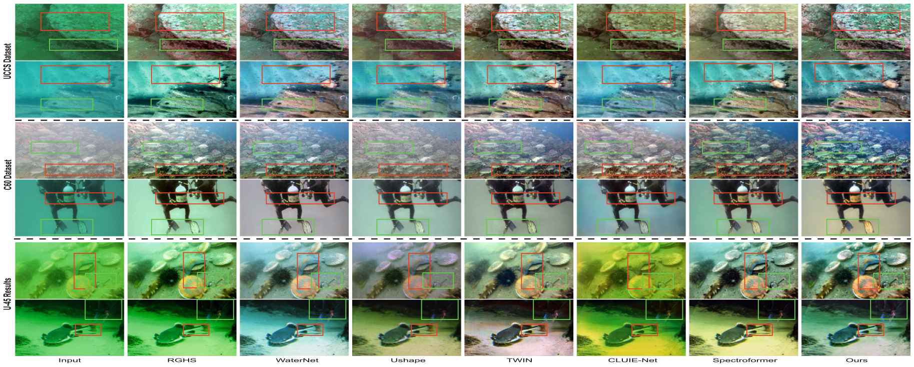


Fig. 3. Qualitative comparison of the proposed method (Ours) with existing methods (RGHS [13], WaterNet[26], CLUIE-Net[29], U-shape[37], TWIN[34], and Sprectroformer[20]) for underwater image enhancement on real-world UCCS, C60 and U45 datasets.

function L_T as a combination of Charbonnier loss L_C [3], Gradient loss L_G [39], Multiscale Structural Similarity Index (MS-SSIM) loss L_M[50], and Perceptual loss L_P.[18]. The effectiveness of the combination of losses is demonstrated in Table-b of Table 3.

4 Experimental Analysis

This section encompasses datasets, training particulars, and comparative analysis of the proposed network.

4.1 Datasets

To perform a comparative analysis, we employed the synthetic Underwater Image Enhancement Benchmark (UIEB), a paired dataset [26], and EUVP[16] dataset includes 11,435 image pairs (clean and degraded) for training and 515 pairs for testing, captured with various cameras and configurations. alongside real-world underwater datasets U45, UCCS, and Challeging 60. Our training dataset comprises 800 randomly selected image pairs, with the remaining 90 images designated for testing. The U45 dataset encompasses 45 real-world images displaying characteristics such as low contrast, color casts, and degradation effects similar to underwater haze. The UCCS dataset comprises 300 genuine underwater images, presenting a diverse array of marine organisms and environments for analysis.

4.2 Training Details

Transformer-based networks necessitate a substantial amount of data samples for effective training, and to enhance the network's generalization capabilities, various data augmentations are applied, including vertical flipping, horizontal flipping, noise addition, and contrast variation. Approximately 4800 images were generated from the UIEB dataset for training purposes, while 90 images from UIEB were allocated for testing. Additionally, we trained our model on the EUVP dataset, which includes 11,435 image pairs (clean and degraded) for training and 515 pairs for testing. To ensure consistency, all training images were resized to a resolution of 256×256. We employed the Adam optimizer with an initial learning rate of 2×10^{-4}. We adjust the learning rate using a scheduler based on the cosine annealing technique. We implemented our model in Pytorch, trained on NVIDIA RTX A4000 16GB GPU.

4.3 Analysis of Synthetic Datasets

The proposed method undergoes a quantitative comparison against existing state-of-the-art techniques, utilizing evaluation metrics such as SSIM, PSNR, and UIQM. Table 1 presents quantitative results for the widely used UIEB and EUVP datasets, while Fig. 2 showcases qualitative results. Our method exhibits competitive performance when compared to state-of-the-art techniques.

4.4 Real-world Dataset Analysis

To evaluate our proposed approach's effectiveness in real-world scenarios, we present results from the U45 and UCCS datasets in Table 2. Our quantitative analysis covers various metrics, including UIQM,UISM, NIQE, and BRISQUE. Summarized results are shown in Table 2. Additionally, qualitative analyses of the U45, UCCS, and C-60 datasets are depicted in Fig. 3. These results highlight significant improvements in color balance and visibility in the enhanced images, attributed to the innovative modules integrated into our proposed method. *More visual results are provided in the supplementary material.*

Table 3. Quantitative results comparison of various network settings (Table-a) and losses settings (Table-b) . *Note: B- Baseline, C- Adaptive Dual-Gated Attentive Fusion Block, D- Adaptive Kernel-based Channel Attention Module, E- Multi Receptive Field Feed Forward Network, F- Color-Attentive Fusion Block module.*

Table-a			Table-b		
Network Setting	PSNR	SSIM	Losses Setting	PSNR	SSIM
B	23.53	0.902	L_C	24.16	0.913
B & C	24.26	0.907	L_C& L_G	24.92	0.920
B & C & D	25.41	0.916	L_C&L_G & L_M	25.17	0.916
B & C & D & E	25.79	0.929	L_C & L_G & L_M & L_P	**26.02**	**0.931**
Ours(B & C & D & E & F)	**26.02**	**0.931**	**Ours**	**26.02**	**0.931**

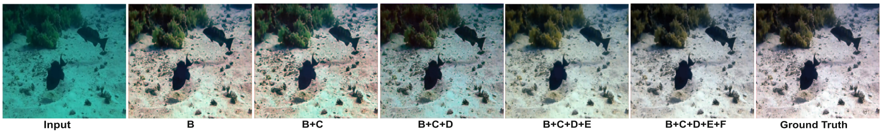


Fig. 4. Qualitative results comparison of various network settings.*Note: B- Baseline, C- Adaptive Dual-Gated Attentive Fusion Block, D- Adaptive kernel-based channel attention, E- Multi Receptive Field Feed Forward Network, F- Color-Attentive Fusion Block module.*

5 Ablation Study

To study the ablation and effectiveness of the proposed blocks, we are using the UIEB [26] dataset.

5.1 Effectiveness of Adaptive Dual-Gated Attentive Fusion Block (ADGAFB)

The Adaptive Dual-Gated Attentive Fusion Block (ADGAFB) combines spatial and frequency domain features to enhance underwater imaging, addressing challenges like scattering, absorption, and structural pattern loss. ADGAFB operates on individual R, G, and B channels, producing color-dependent enhancements. Experimental results confirm its effectiveness, as seen in Table 3.

5.2 Effectiveness of Color-Attentive Fusion Block

This block maintains color fidelity by providing color-wise attention from each R, G, and B channel feature with the incoming original RGB features, effectively addressing challenges related to color distortion and contrast reduction. This operation achieves a natural global enhancement of color and contrast in degraded image features. Experimental results confirm its effectiveness, as shown in Table 3 and Fig. 4.

5.3 Effectiveness of Multi Receptive Field Forward Network

Within the Transformer architecture, the Multi-Receptive Field Forward Network (MRFFN) efficiently forwards attentive features of degraded images, both globally and locally, helps in the deep reconstruction of distinct objects. Experimental results confirm its effectiveness, as shown in Table 3 and Fig. 4.

5.4 Effectiveness of Adaptive Kernel-based Channel Attention Module

The Adaptive Kernel-Based Channel Attention Module enhances feature transmission between the encoder and decoder, improving performance and feature augmentation. Experimental results confirm its effectiveness, as shown in Table 3 and Fig. 4.

5.5 Effectiveness of various loss functions

We have evaluated the impact of different loss functions, including Charbonnier loss (L_C), Gradient loss (L_G), Multiscale Structural Similarity Index (MS-SSIM) loss (L_M), and Perceptual loss (L_P). As shown in Table-b of Table 3, these experiments have helped us verify the training stability and robustness of our proposed method.

6 Downstream Application of Our Proposed Model (ACFTNet)

Our proposed ACFTNet demonstrates versatility across multiple downstream applications, such as image segmentation, underwater monocular depth estimation, and saliency detection. Specifically, we highlight saliency detection, which involves identifying the most significant regions within images, as depicted in Fig. 5.

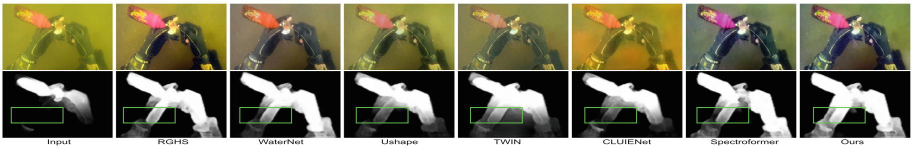


Fig. 5. Qualitative comparison for saliency detection map of the outputs of proposed method (Ours) with existing methods.

7 Limitation and Conclusion

7.1 Limitation

The current methods encounter significant difficulties in enhancing muddy and blurry underwater images. Our approach consistently surpasses state-of-the-art techniques, demonstrating superior performance. However, deblurring remains an area requiring further improvement, which will be a primary focus for our future work.

7.2 Conclusion

In this paper, we proposed ACFTNet: Attentive Color Fusion Transformer Network, an underwater image enhancement model designed as a supervised method featuring several key components. Central to ACFTNet is the Adaptive Dual Gated Attentive Feature Fusion block, which effectively extracts and integrates localized and global illumination features from both spatial and frequency domains. Additionally, we introduce a Color-Attentive Feature Fusion block aimed at enhancing individual R, G, and B color channels and correlating these features with the original RGB image to improve color efficacy. Furthermore, the Multi-Receptive Field Feed Forward Network facilitates the propagation of diverse deep features, offering detailed preservation of edges and textures for superior reconstruction. Extensive testing and analysis were conducted on various datasets, including real, synthetic, paired, and unpaired datasets, demonstrating the effectiveness of our model through quantitative and qualitative results. Moreover, our model exhibits versatility and applicability across different applications.

Acknowledgement. This work was made possible through the generous support of the projects MoES/PAMC/DOM/04/2022 (E-12710) and TIHIITG202204. The authors would also like to thank all the members of CVPR Lab for their support on this work.

References

1. Abdul Ghani, A.S., Mat Isa, N.A.: Underwater image quality enhancement through composition of dual-intensity images and rayleigh-stretching. Springerplus **3**, 1–14 (2014)
2. Ancuti, C., Ancuti, C.O., Haber, T., Bekaert, P.: Enhancing underwater images and videos by fusion. In: 2012 IEEE conference on computer vision and pattern recognition. pp. 81–88. IEEE (2012)
3. Bruhn, A., Weickert, J., Schnörr, C.: Lucas/kanade meets horn/schunck: Combining local and global optic flow methods. Int. J. Comput. Vision **61**, 211–231 (2005)
4. Chiang, J.Y., Chen, Y.C.: Underwater image enhancement by wavelength compensation and dehazing. IEEE Trans. Image Process. **21**(4), 1756–1769 (2011)
5. Drews, P., Nascimento, E., Moraes, F., Botelho, S., Campos, M.: Transmission estimation in underwater single images. In: Proceedings of the IEEE international conference on Computer Vision Workshops. pp. 825–830 (2013)
6. Drews, P.L., Nascimento, E.R., Botelho, S.S., Campos, M.F.M.: Underwater depth estimation and image restoration based on single images. IEEE Comput. Graphics Appl. **36**(2), 24–35 (2016)
7. Du, D., Li, E., Si, L., Xu, F., Niu, J., Sun, F.: Uiedp: Underwater image enhancement with diffusion prior. arXiv preprint arXiv:2312.06240 (2023)
8. Dudhane, A., Hambarde, P., Patil, P., Murala, S.: Deep underwater image restoration and beyond. IEEE Signal Process. Lett. **27**, 675–679 (2020)
9. Fabbri, C., Islam, M.J., Sattar, J.: Enhancing underwater imagery using generative adversarial networks. In: 2018 IEEE international conference on robotics and automation (ICRA). pp. 7159–7165. IEEE (2018)
10. He, K., Sun, J., Tang, X.: Single image haze removal using dark channel prior. IEEE Trans. Pattern Anal. Mach. Intell. **33**(12), 2341–2353 (2010)
11. Henderson, J., Pizarro, O., Johnson-Roberson, M., Mahon, I.: Mapping submerged archaeological sites using stereo-vision photogrammetry. Int. J. Naut. Archaeol. **42**(2), 243–256 (2013)
12. Hou, G., Li, J., Wang, G., Yang, H., Huang, B., Pan, Z.: A novel dark channel prior guided variational framework for underwater image restoration. J. Vis. Commun. Image Represent. **66**, 102732 (2020)

13. Huang, D., Wang, Y., Song, W., Sequeira, J., Mavromatis, S.: Shallow-water image enhancement using relative global histogram stretching based on adaptive parameter acquisition. In: MultiMedia Modeling: 24th International Conference, MMM 2018, Bangkok, Thailand, February 5-7, 2018, Proceedings, Part I 24. pp. 453–465. Springer (2018)
14. Iqbal, K., Odetayo, M., James, A., Salam, R.A., Talib, A.Z.H.: Enhancing the low quality images using unsupervised colour correction method. In: 2010 IEEE International Conference on Systems, Man and Cybernetics. pp. 1703–1709. IEEE (2010)
15. Islam, M.J., Luo, P., Sattar, J.: Simultaneous enhancement and super-resolution of underwater imagery for improved visual perception. arXiv preprint arXiv:2002.01155 (2020)
16. Islam, M.J., Xia, Y., Sattar, J.: Fast underwater image enhancement for improved visual perception. IEEE Robotics and Automation Letters (RA-L) **5**(2), 3227–3234 (2020)
17. Jiang, Z., Li, Z., Yang, S., Fan, X., Liu, R.: Target oriented perceptual adversarial fusion network for underwater image enhancement. IEEE Trans. Circuits Syst. Video Technol. **32**(10), 6584–6598 (2022)
18. Johnson, J., Alahi, A., Fei-Fei, L.: Perceptual losses for real-time style transfer and super-resolution. In: Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II 14. pp. 694–711. Springer (2016)
19. Khan, M.R., Kulkarni, A., Phutke, S.S., Murala, S.: Underwater image enhancement with phase transfer and attention. In: 2023 International Joint Conference on Neural Networks (IJCNN). pp. 1–8. IEEE (2023)
20. Khan, R., Mishra, P., Mehta, N., Phutke, S.S., Vipparthi, S.K., Nandi, S., Murala, S.: Spectroformer: Multi-domain query cascaded transformer network for underwater image enhancement. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 1454–1463 (2024)
21. Kim, H.G., Seo, J.M., Kim, S.M.: Comparison of gan deep learning methods for underwater optical image enhancement. Journal of Ocean Engineering and Technology **36**(1), 32–40 (2022)
22. Kong, L., Dong, J., Ge, J., Li, M., Pan, J.: Efficient frequency domain-based transformers for high-quality image deblurring. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5886–5895 (2023)
23. Li, C.Y., Guo, J.C., Cong, R.M., Pang, Y.W., Wang, B.: Underwater image enhancement by dehazing with minimum information loss and histogram distribution prior. IEEE Trans. Image Process. **25**(12), 5664–5677 (2016)
24. Li, C., Anwar, S., Hou, J., Cong, R., Guo, C., Ren, W.: Underwater image enhancement via medium transmission-guided multi-color space embedding. IEEE Trans. Image Process. **30**, 4985–5000 (2021)
25. Li, C., Anwar, S., Porikli, F.: Underwater scene prior inspired deep underwater image and video enhancement. Pattern Recogn. **98**, 107038 (2020)
26. Li, C., Guo, C., Ren, W., Cong, R., Hou, J., Kwong, S., Tao, D.: An underwater image enhancement benchmark dataset and beyond. IEEE Trans. Image Process. **29**, 4376–4389 (2019)
27. Li, C., Guo, J., Guo, C.: Emerging from water: Underwater image color correction based on weakly supervised color transfer. IEEE Signal Process. Lett. **25**(3), 323–327 (2018)

28. Li, J., Skinner, K.A., Eustice, R.M., Johnson-Roberson, M.: Watergan: Unsupervised generative network to enable real-time color correction of monocular underwater images. IEEE Robotics and Automation letters **3**(1), 387–394 (2017)
29. Li, K., Wu, L., Qi, Q., Liu, W., Gao, X., Zhou, L., Song, D.: Beyond single reference for training: Underwater image enhancement via comparative learning. IEEE Transactions on Circuits and Systems for Video Technology (2022)
30. LI, T.H., YU, Z.H., YU, Z.D.: Dual-branch low-light image enhancement network combined with spatial and frequency domain information. Journal of Computer Applications p. 0 (2023)
31. Liang, Z., Ding, X., Wang, Y., Yan, X., Fu, X.: Gudcp: Generalization of underwater dark channel prior for underwater image restoration. IEEE Trans. Circuits Syst. Video Technol. **32**(7), 4879–4884 (2021)
32. Lim, L.A., Keles, H.Y.: Learning multi-scale features for foreground segmentation. Pattern Anal. Appl. **23**(3), 1369–1380 (2020)
33. Liu, C., Jia, S., Wu, H., Zeng, D., Cheng, F., Zhang, S.: A spatial-frequency domain associated image-optimization method for illumination-robust image matching. Sensors **20**(22), 6489 (2020)
34. Liu, R., Jiang, Z., Yang, S., Fan, X.: Twin adversarial contrastive learning for underwater image enhancement and beyond. IEEE Trans. Image Process. **31**, 4922–4936 (2022)
35. Mao, Y., Chen, K., Diao, W., Sun, X., Lu, X., Fu, K., Weinmann, M.: Beyond single receptive field: A receptive field fusion-and-stratification network for airborne laser scanning point cloud classification. ISPRS J. Photogramm. Remote. Sens. **188**, 45–61 (2022)
36. Pang, X., Yin, Y., Zheng, Y.: Multi-receptive field soft attention part learning for vehicle re-identification. Entropy **25**, 594 (03 2023). https://doi.org/10.3390/e25040594 <error l="305" c="Invalid command: paragraph not started." />
37. Peng, L., Zhu, C., Bian, L.: U-shape transformer for underwater image enhancement. IEEE Transactions on Image Processing (2023)
38. Peng, Y.T., Cao, K., Cosman, P.C.: Generalization of the dark channel prior for single image restoration. IEEE Trans. Image Process. **27**(6), 2856–2868 (2018)
39. Ribeiro, J., Elsayed, E.: A case study on process optimization using the gradient loss function. Int. J. Prod. Res. **33**(12), 3233–3248 (1995)
40. Schechner, Y., Karpel, N.: Clear underwater vision. vol. 1, pp. I–536 (01 2004). https://doi.org/10.1109/CVPR.2004.1315078
41. Schettini, R., Corchs, S.: Underwater image processing: state of the art of restoration and image enhancement methods. EURASIP journal on advances in signal processing **2010**, 1–14 (2010)
42. Sharma, P., Bisht, I., Sur, A.: Wavelength-based attributed deep neural network for underwater image restoration. ACM Trans. Multimed. Comput. Commun. Appl. **19**(1), 1–23 (2023)
43. Shi, W., Caballero, J., Huszár, F., Totz, J., Aitken, A.P., Bishop, R., Rueckert, D., Wang, Z.: Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1874–1883 (2016)
44. Singh, G., Mittal, A.: Various image enhancement techniques-a critical review. International Journal of Innovation and Scientific Research **10**(2), 267–274 (2014)
45. Song, J., Meng, C., Ermon, S.: Denoising diffusion implicit models. arXiv preprint arXiv:2010.02502 (2020)

46. Syariz, M.A., Lin, C.H., Nguyen, M.V., Jaelani, L.M., Blanco, A.C.: Waternet: A convolutional neural network for chlorophyll-a concentration retrieval. Remote Sensing **12**(12), 1966 (2020)
47. Treibitz, T., Schechner, Y.Y.: Active polarization descattering. IEEE Trans. Pattern Anal. Mach. Intell. **31**(3), 385–399 (2008)
48. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in neural information processing systems **30** (2017)
49. Wang, D., Sun, Z.: Frequency domain based learning with transformer for underwater image restoration. In: Pacific Rim International Conference on Artificial Intelligence. pp. 218–232. Springer (2022)
50. Wang, Z., Simoncelli, E.P., Bovik, A.C.: Multiscale structural similarity for image quality assessment. In: The Thrity-Seventh Asilomar Conference on Signals, Systems & Computers, 2003. vol. 2, pp. 1398–1402. Ieee (2003)
51. Williams, D.P.: On optimal auv track-spacing for underwater mine detection. In: 2010 IEEE International Conference on Robotics and Automation. pp. 4755–4762. IEEE (2010)
52. Yang, H.Y., Chen, P.Y., Huang, C.C., Zhuang, Y.Z., Shiau, Y.H.: Low complexity underwater image enhancement based on dark channel prior. In: 2011 Second International Conference on Innovations in Bio-inspired Computing and Applications. pp. 17–20 (2011). https://doi.org/10.1109/IBICA.2011.9
53. Yang, M., Hu, K., Du, Y., Wei, Z., Sheng, Z., Hu, J.: Underwater image enhancement based on conditional generative adversarial network. Signal Processing: Image Communication **81**, 115723 (2020)
54. Zamir, S.W., Arora, A., Khan, S., Hayat, M., Khan, F.S., Yang, M.H.: Restormer: Efficient transformer for high-resolution image restoration. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 5728–5739 (2022)
55. Zhang, H., Cisse, M., Dauphin, Y.N., Lopez-Paz, D.: mixup: Beyond empirical risk minimization. arXiv preprint arXiv:1710.09412 (2017)

Unsupervised Low-Light Image Enhancement with Dual Contrastive Learning

Yujie Wang, Bing Li, Jie Huang, and Feng Zhao(✉)

University of Science and Technology of China, Hefei 230027, China
wangyj1021@mail.ustc.edu.cn, bing0123@mail.ustc.edu.cn,
hj0117@mail.ustc.edu.cn, fzhao956@ustc.edu.cn

Abstract. Conventional learning-based approaches for enhancing low-light images typically rely on a large amount of paired training data, which is challenging to obtain in practice. To address this issue, some algorithms have been proposed using the generative adversarial mechanism to utilize unpaired data. However, these methods commonly utilize perceptual constraints to preserve content information, leading to incorrect lightness enhancement. In this paper, we propose a dual contrastive learning scheme for unsupervised low-light image enhancement (LLIE), aiming to balance the lightness enhancement and content preservation. Specifically, we introduce two models with distinct biases towards lightness enhancement and content preservation, respectively. These models produce intermediate results that serve as negative and positive samples, guiding the final model to generate the desired outcome. Considering the coupling of luminance and noise in low-light conditions, we propose a Frequency-Spatial Attention Module to obtain an adaptive illumination map to guide light enhancement and noise removal. Extensive experimental results demonstrate our superiority over several state-of-the-art methods.

Keywords: Low-light image enhancement · Contrastive learning · Generative adversarial networks

1 Introduction

Images captured in low-light conditions commonly exhibit issues such as poor visibility, color inaccuracies, and acute noise, leading to unfavorable visual perception and hindering human annotation. Consequently, researchers have proposed several low-light image enhancement methods over the past decades. These methods aim to enhance visibility, mitigate noise, and improve overall visual quality. As a result, they have proven beneficial for various downstream computer vision tasks, including object detection and action recognition.

Y. Wang and B. Li—Both authors contributed equally to this research.

A. Antonacopoulos et al. (Eds.): ICPR 2024, LNCS 15321, pp. 325–338, 2025.
https://doi.org/10.1007/978-3-031-78305-0_21

Traditional approaches such as classical histogram equalization (HE)[11] and Gamma Correction (GC) methods focus on compressing the dynamic range and non-linearly adjusting the input to enhance image contrast. However, these methods overlook the interconnections among pixels and the preservation of naturalness in the enhanced results. Another research direction involves leveraging the Retinex theory [14] to model the process of decomposing input images into separate components: illumination and reflectance. This approach considers the reflectance component as the enhanced output or generates enhanced results by manipulating the illumination component [8,25].

Nevertheless, a common issue encountered in enhancement methods is the tendency to amplify latent noise. To address this concern, several approaches have been proposed to mitigate the residual noise present in the reflectance component [17,39]. However, these methods have exhibited limited effectiveness due to the spatially-varying nature of noise measurement.

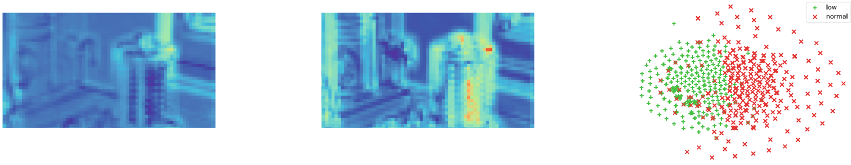
Recently, data-driven low-light enhancement techniques have achieved significant progress under the rapid development of deep learning, demonstrating markedly superior performance compared to traditional methods[1–4,13,18,19,32,35,37–39]. A kind of learning-based approaches follow the Retinex theory that decomposes the low-light image into illumination and reflectance, then restore them respectively in a data-driven manner [1,3,35,37–39]. Another research direction involves the acquisition of diverse frequency representations followed by progressive recovery[32]. However, these methods require a large amount of paired data, which is difficult to obtain in the real-world scenario, limiting their wide-range application.

To eliminate the reliance on paired data, unsupervised learning methods[5,7,12,20,23,39,40] have been developed for image enhancement.While unsupervised learning methods employ generative adversarial mechanism [6] to encourage the distribution of the enhanced image to be close to the target normal-light image [12,28,34], EnlightenGAN [12] employs a global-local discriminator structure and self-regularization loss that assist the generator in enhancing lightness.With the employ of the GAN mechanism, these approaches can generate images with a similar overall appearance with normal-light images in some cases. Nevertheless, GAN mechanism usually generates results with content distortion [31]. To address this issue, EnlightenGAN [12] employ the perceptual constraint between the input and output to ensure their consistent content, and this constraint is usually implemented by the pre-trained classifier. However, as depicted in Fig. 1, the features of the low-light images in the pre-trained VGG [27] classifier are significantly different from those of normal-light images. Based on this, the perceptual constraint often misleads the enhancement model to learn the degradation of the low-light input, resulting in generating lightness distortion in the enhanced result.

To break this limitation, we propose a dual contrastive learning scheme to assist an unpaired learning strategy for low-light image enhancement. As shown in Fig. 3, we find that the model trained dominantly by the GAN mechanism generates images with correct lightness but corrupted content, while the model trained dominantly by the perceptual constraint produces images with preserved content but distorted lightness. These two kinds of results are complementary to each other and thus can guide the enhancement of the low-light image. To this end, we decouple the whole process of unpaired learning into two stages. In the first stage, we propose learning two models mainly driven by the GAN mechanism and the perceptual constraint, respectively. While in the second stage, we train a model driven by the contrastive term and GAN mechanism. Specifically, the two types of intermediate results produced during the initial stage function as reciprocal positive and negative samples, which guide the enhancement model to learn correct lightness and content representations in a pixel constraint way as shown in Fig. 2. In addition, current approaches commonly utilize the maximum channel of the image as the illumination map, which neglects the inherent stochastic degradation and the coupling of illumination enhancement and degradation present in the image. To tackle the issue, we propose a Frequency-Spatial Attention Module (FSAM) that enables the model to acquire an adaptive illumination map to guide illumination enhancement and noise removal.

In summary, our primary contributions are:

- We perform a comprehensive analysis of previous unpaired learning methods which reveals the deficiencies of perceptual loss in low-light image enhancement and propose a dual contrastive learning scheme to enhance the effectiveness of unpaired low-light image enhancement.
- We propose to train two models, each with a specific bias towards lightness enhancement and content preservation. Their outputs are categorized as positive and negative samples relative to each other, thereby guiding the enhancement process of the final enhancement model.
- We introduce a Frequency-Spatial Attention Module to acquire an adaptive illumination map to guide lightness enhancement and noise removal.
- Extensive experiments are conducted to illustrate the superiority of our method against other state-of-the-art approaches.



(a) (b) (c)

Fig. 1. (a) and (b) are the feature maps of low-light image and normal-light image on the pre-trained VGG extractor, respectively. (c) The feature distribution of (a) and (b). The feature distribution of the low-light image is significantly different from that of normal-light image.

2 Method

2.1 Preliminary Analysis

We provide feature maps visualization and a TSNE analysis in Fig. 1 to analyze the statistical distribution of enhancement results and realistic normal-light targets.

As shown in Fig. 1, low-light and normal light features extracted by the pre-trained VGG model have different attentions, where specific objects and overall texture are focused, respectively. In addition, the TSNE analysis of low-light and normal-light features distribution also indicates that though extracting features with pre-trained VGG models reduces the domain gap, it is still quite easy to distinguish different domains. Given the aforementioned analysis, we draw the conclusion that performing constraints between low-light inputs and normal-light outputs with VGG network to preserve texture information is not reasonable.

We further observe that the outcomes produced by the GAN mechanism-dominated model and the perceptual-dominated method exhibit complementary characteristics in terms of brightness and texture. Consequently, we propose employing contrastive learning to address these disparities, as detailed in Sec. 2.4.

2.2 Overview

As shown in Fig. 2, our method consists of two phases: the **Sample Preparation Stage** and the **Refinement Stage** with generators (G_{GAN}, G_{Per} and G_{Ref}) of the same architecture. The Samples Preparing Stage first trains G_{GAN} and G_{Per} dominated by GAN mechanism and perceptual constraint, respectively. The Refinement Stage then takes advantage of the complementary models from the first stage and constructs dual contrastive terms to refine illumination and achieve perceptual satisfaction for G_{Ref}.

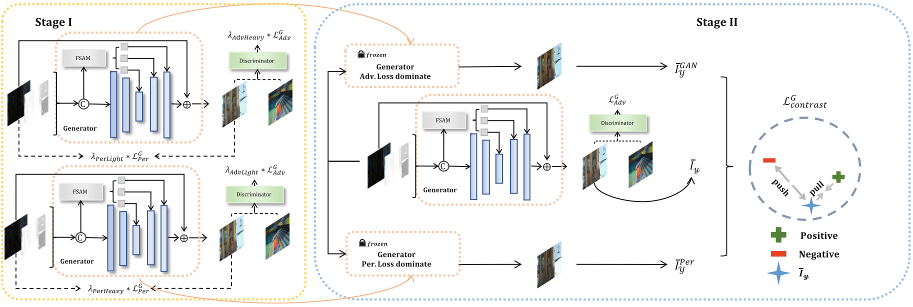


Fig. 2. The overall architecture of our generator and dual contrastive method. Stage I is the samples preparing stage and Stage II is the refinement stage (see Section 2).

2.3 Stage I: Sample Preparation Stage

We briefly describe our generator architecture as follows. As illustrated in Fig. 2, our generator is built based on U-Net with skip connections modulated by an illumination guidance block. Given a low-light input I_{low}, we concatenate low-light I_{low} with the corresponding illumination map I_{illu} which is estimated by seeking the maximum value of the RGB three color channels for each individual pixel[8], together forming the initial inputs $[I_{low},\ I_{illu}]$.Then we apply the Frequency-Spatial Attention Module (FSAM) to get an adaptive illumination map which is propagated to the subsequent layers for further processing. After that we combine the low-light input with the network output via a skip connection, and the final result is the prepared sample.

To achieve two models with distinct biases, we assign varying loss weights to the adversarial and perceptual components. Specifically, the lightness-accurate samples generator is obtained by setting the loss weights of the adversarial and perceptual as $\lambda_{AdvHeavy}$ and $\lambda_{PerLight}$, respectively. Conversely, the content-preserving sample generator is obtained by setting the loss weights of the adversarial and perceptual as $\lambda_{AdvLight}$ and $\lambda_{PerHeavy}$.

Due to the coupling of luminance and noise in low-light conditions [29], it is significantly challenging to address both noise removal and light enhancement simultaneously. Prevailing enhancement methods commonly adopt a direct approach by utilizing the max-channel as the illumination map to guide light enhancement. This approximation can be easily influenced by color information, which results in a deterministic mapping as it overlooks the stochastic nature of noise.

Through the analysis of the relationship between noise and lightness[32], regions characterized by inadequate lightness tend to exhibit higher levels of noise, while regions with sufficient lightness typically have lower noise levels. Addressing regions with insufficient light and higher noise necessitates long-range (global) operations for restoration, whereas regions with lower noise and adequate lightness favor short-range (local) operations. Previous studies [10,16] have demonstrated that the Fourier Transform can partially decompose luminance and noise in the Fourier domain, with luminance information predominantly represented as amplitudes and noise manifested in phases. It inspires us to implement phase enhancement and amplitude enhancement in parallel in the Fourier domain, and feature enhancement in the spatial domain. Furthermore, the Fourier Transform possesses the capability to extract global information without imposing an excessive number of parameters on the neural network. Motivated by these observations, we propose a Frequency-Spatial Attention Module (FSAM) to modulate the max channel map, enabling the acquisition of an adaptive illumination map that guides both luminance enhancement and noise removal. Detailed elaboration on the FSAM will be provided in section 2.5.

To adapt to current research conditions and achieve unpaired learning, a generative adversarial strategy is adopted in the process of training, and we keep the same discriminator design with EnlightenGAN [12].

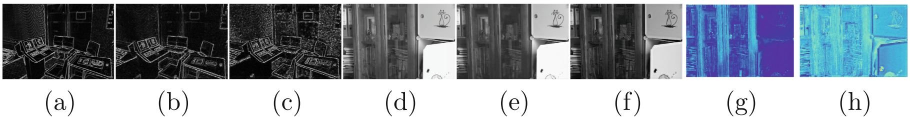


Fig. 3. We employ the results of the Prewitt operator to measure the texture of the image. (a), (b), and (c) are the results of the Prewitt operator applied to normal-light image, and outputs of the generator dominated by GAN mechanism and perceptual constraint, respectively. As for the illumination measurement, we first decompose the images into a luminance component (Y) and a chromatic component (UV). Then, we employ the luminance component to measure the illumination of the image. (d), (e), and (f) are the luminance components of the normal-light image, and outputs of the generator dominated by the GAN mechanism and perceptual constraint, respectively. (g) is the absolute value of the pixel-by-pixel difference between (d) and (e), while (h) is the absolute value of the pixel-by-pixel difference between (d) and (f).

2.4 Stage II: Refinement Stage

After obtaining results from Stage I, it is worth noting that results generated by GAN mechanism-dominated model is well enlightened but with corrupted content, while the model trained dominantly by the perceptual constraint produces images with distorted lightness and preserved content. We fully take advantage of both GAN mechanism-driven model and perceptual constraint-driven models, thus proposing the dual contrastive term.

For texture contrastive term $\mathcal{L}^G_{texture}$, we still apply VGG to extract component of high-frequency and state as

$$\mathcal{L}^G_{texture} = \frac{||\phi(\tilde{I}^{Per}_y) - \phi(\tilde{I}_y)||_1}{||\phi(\tilde{I}^{GAN}_y) - \phi(\tilde{I}_y)||_1}, \tag{1}$$

where $\tilde{I}^{GAN}_y$ and $\tilde{I}^{Per}_y$ denote the outputs of GAN mechanism-dominated model and perceptual constraint-dominated model, respectively.

We define lightness contrastive term $\mathcal{L}^G_{illu}$ as

$$\mathcal{L}^G_{illu} = ||1 - \frac{l(\tilde{I}^{GAN}_y, \tilde{I}_y)}{l(\tilde{I}^{Per}_y, \tilde{I}_y)}||_1, \tag{2}$$

with definition of $l(\mathbf{x}, \mathbf{y}) = \frac{2\mu_x\mu_y + C_1}{\mu_x^2 + \mu_y^2 + C_1}$ from SSIM loss, $C_1 = (K_1 L)^2$. Small constant $K_1 \ll 1$. L, μ_x and μ_y are the dynamic range of the pixel values, mean intensity of image x and mean intensity of image y, respectively.

Dual contrastive loss $\mathcal{L}^G_{contrast}$ is therefore expressed as

$$\mathcal{L}^G_{contrast} = \lambda_1 \mathcal{L}^G_{texture} + \lambda_2 \mathcal{L}^G_{illu}. \tag{3}$$

To summarize, total loss function for Generator in stage II is as follows:

$$\mathcal{L}^G_{total} = \mathcal{L}^G_{Adv} + \lambda_3 \mathcal{L}^G_{contrast}, \tag{4}$$

where we keep the adversarial term $\mathcal{L}_{Adv}^{G}$ for unpaired learning and naturalness preservation. In this work, λ_1,λ_2,λ_3 are empirically set to 2.5,0.5,1, respectively.

2.5 Frequency-Spatial Attention Module

Fourier Transform Given an image input $x \in R^{H\times W\times C}$, we can convert it to the Fourier space by using The Fourier Transform, which can be formulated as follows:

$$F(X)(u,v) = \frac{1}{\sqrt{HW}} \sum_{h=0}^{H-1} \sum_{w=0}^{W-1} x(h,w) e^{-j2\pi(\frac{h}{H}u + \frac{w}{W}v)}. \tag{5}$$

We can get the amplitude component A(x)(u,v) and phase P(x)(u,v) by the equation:

$$A(x)(u,v) = \sqrt{R^2(x)(u,v) + I^2(x)(u,v)}, \tag{6}$$

$$P(x)(u,v) = arctan[\frac{I(x)(u,v)}{R(x)(u,v)}], \tag{7}$$

where $R(x)(u,v)$ and $I(x)(u,v)$ denote the real and imaginary part, respectively. **Frequency-Spatial Attention Module** As shown in Fig. 4, the input concatenated features are bifurcated into two distinct branches: the frequency branch and the spatial branch, which are depicted on the left and right sides, respectively. We apply the Fast Fourier Transform (FFT) on the frequency branch to decompose the luminance and noise to a certain extent in the Fourier domain and obtain the amplitude and phase components. To process luminance and noise separately and extract global information, we apply two 1 × 1 convolutional layers with a Leaky-ReLU activation separately to the amplitude and phase components. Finally, we transform them back to the spatial domain by Inverse FFT to fuse them with the results obtained by the spatial branch, and a 3 × 3 convolutional layer is applied to stabilize the training of the FSAM.

As a spatial local detail complementary branch, the spatial branch uses the 3 × 3 convolution and ReLU to effectively model the structural dependency in the spatial domain.

Given the input $X_{in} = contact(I_{low}, I_{illu})$, the frequency branch $F(X_{in})$ and spatial branch $S(X_{in})$, the final output feature $I'_{illu} \in R^{C_{illu}\times H\times W}$ by FSAM can be obtained as follows:

$$I'_{illu} = I_{illu} \otimes M(X_{in}) = I_{illu} \otimes \sigma(F(X_{in}) + S(X_{in})), \tag{8}$$

where $M(X_{in}) \in R^{C_{illu}\times H\times W}$ represent the attentional map generated by FSAM. $\otimes$ represents the element-wise multiplication.

3 Experiments

Dataset and Evaluation. We perform experiments on unpaired no-reference datasets which include LIME [8], MEF [22], NPE [30], DICM [15], and paired LOL dataset [33] but with unpaired learning manner. The LOL dataset [33]

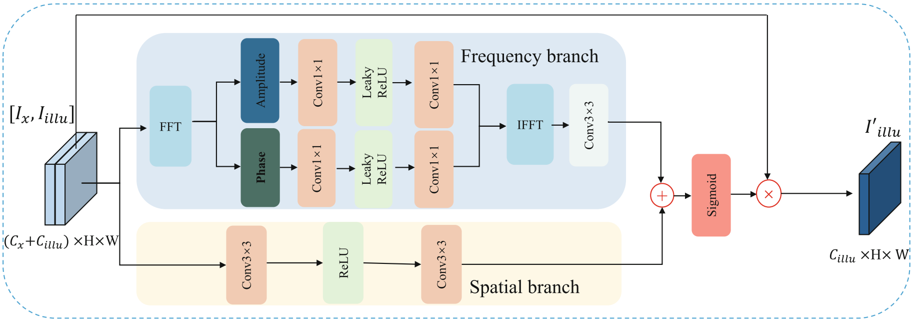
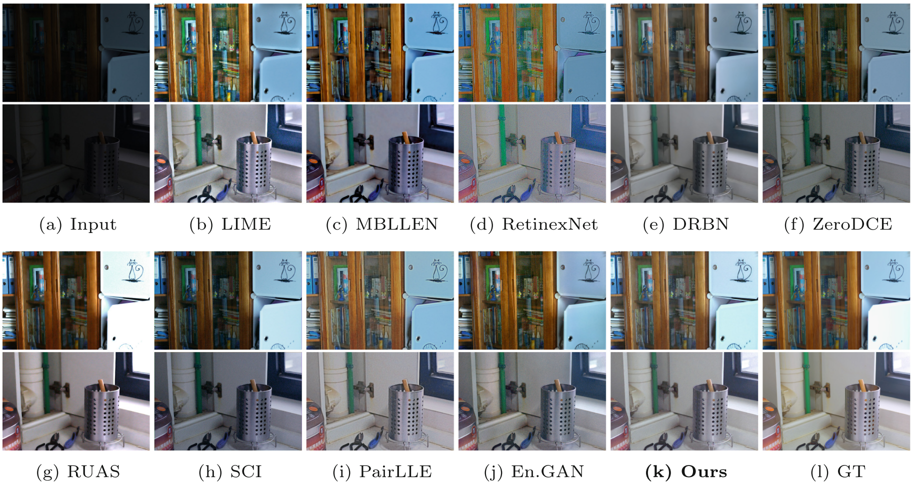


Fig. 4. The Frequency-Spatial Attention Module.


(a) Input (b) LIME (c) MBLLEN (d) RetinexNet (e) DRBN (f) ZeroDCE

(g) RUAS (h) SCI (i) PairLLE (j) En.GAN **(k) Ours** (l) GT

Fig. 5. Qualitative results on the LOL dataset. En.GAN denotes EnlightenGAN. Please zoom in for details.

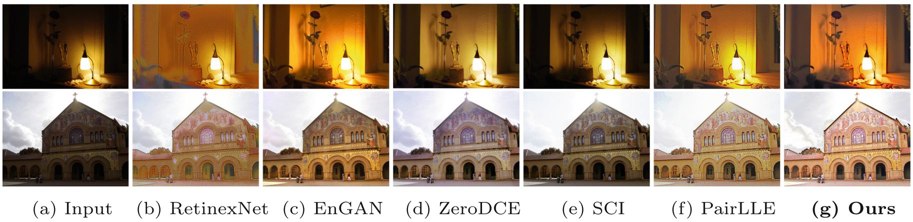
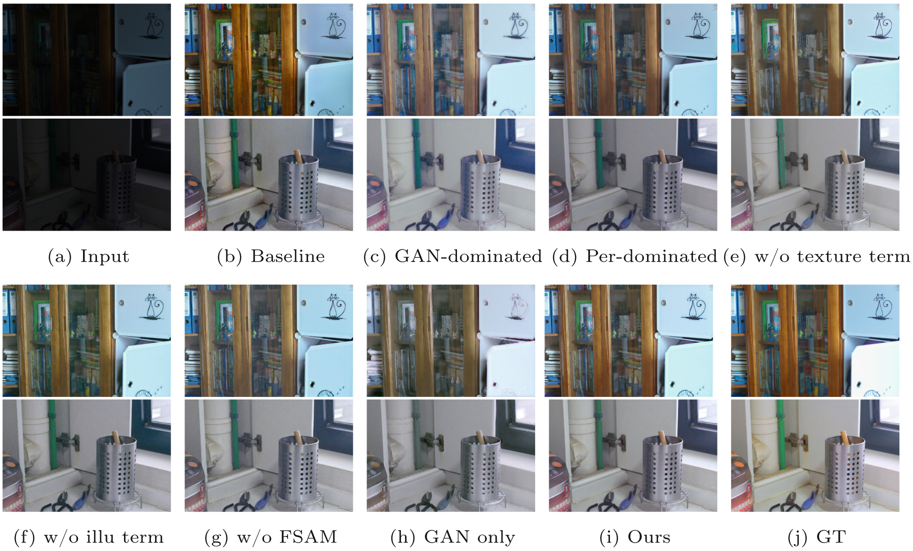

(a) Input (b) RetinexNet (c) EnGAN (d) ZeroDCE (e) SCI (f) PairLLE **(g) Ours**

Fig. 6. Qualitative results on the no-reference dataset, where the first row is the results on the MEF dataset, and the second row is the results on the DICM dataset. En.GAN denotes EnlightenGAN. Please zoom in for details.

Table 1. Ablation studies on the **LOL** dataset.

Type	PSNR ↑	SSIM ↑	MS-SSIM ↑	LPIPS ↓	FID ↓	NIQE ↓
Baseline	18.76	0.731	0.835	0.2416	93.21	4.2478
GAN-dominated	18.96	0.750	0.893	0.1733	96.68	4.3012
Perceptual-dominated	19.09	0.733	0.891	0.1958	88.40	4.1843
w/o texture term	19.15	0.729	0.871	0.1693	113.54	4.8218
w/o illu term	19.63	0.733	0.902	0.1805	88.66	4.0659
w/o FSAM	19.42	0.732	0.894	0.1816	92.55	4.2878
GAN only	18.25	0.685	0.835	0.2322	161.74	4.2822
Ours	**20.27**	**0.754**	**0.905**	**0.1552**	**78.31**	**3.9435**


(a) Input (b) Baseline (c) GAN-dominated (d) Per-dominated (e) w/o texture term

(f) w/o illu term (g) w/o FSAM (h) GAN only (i) Ours (j) GT

Fig. 7. Qualitative results of different settings of the ablation studies in Table. 1. "Per-dominated" denotes perceptual-dominated. Please zoom in for details.

consists of 485 pairs for training and 15 pairs for testing, we randomly shuffled images to guarantee unpaired learning.

For evaluation with the ground truth (GT), we apply the following metrics: PSNR, SSIM, MS-SSIM and LPIPS [36]. In addition, we adopt non-reference image quality metrics including FID [9] and NIQE [24](lower is better).

Implementation Details. During training, we randomly crop input images into 256x256. and the whole training process takes total 200 epochs and we optimize our model with Adam optimizer, of which the initial learning rate is set as 1e-4 and decreases linearly from the 100th epoch. We implement our framework with PyTorch and all experiments are performed on a single GPU of 3080Ti.

Table 2. Quantitative results on the **LOL dataset**. "T", "S", and "U" represent "Traditional", "Supervised", and "Unsupervised" methods, respectively. En.GAN denotes EnlightenGAN.

Type	Method	PSNR ↑	SSIM ↑	MS-SSIM ↑	LPIPS ↓	FID ↓	NIQE ↓
T	HE[11]	**14.54**	0.377	0.640	0.5036	118.39	8.437
	CLAHE [26]	9.83	0.397	0.510	0.4225	107.37	**7.394**
	LIME[8]	14.22	**0.514**	**0.767**	**0.3683**	**97.50**	8.058
U	CycleGAN[40]	19.51	0.746	0.860	0.2377	102.66	**3.4049**
	En.GAN[12]	18.76	0.731	0.835	0.2416	93.21	4.2478
	RRDNet[39]	11.00	0.440	0.601	0.3710	92.78	7.4306
	ZeroDCE[7]	16.24	0.511	0.729	0.4012	140.45	7.8830
	RUAS[20]	16.40	0.500	0.822	0.2701	112.40	6.3418
	SCI[23]	14.78	0.522	0.854	0.3393	93.21	7.8726
	PairLIE[5]	19.51	0.736	0.891	0.2477	87.19	4.0847
	Ours	**20.27**	**0.754**	**0.905**	**0.1552**	**78.31**	3.9435
S	RetinexNet[33]	16.54	0.709	0.672	0.3179	184.48	5.5463
	MBLLEN[21]	**18.98**	0.816	0.866	0.1400	**67.38**	**4.2301**
	DRBN[34]	18.80	**0.830**	**0.931**	**0.1009**	74.55	5.1131

3.1 Comparison with state of the art

Quantitative Comparison. Tables 2 and 3 provide quantitative comparisons with other methods on the LOL dataset and no-reference datasets(MEF, LIME, NPE, and DICM), respectively. Table. 4 reports the model parameters, FLOPs and runtime averaged over 50 images of size 512x512. As can be seen, our method achieves the best performance in terms of most metrics on the LOL dataset and the no-reference datasets(MEF, LIME, NPE, and DICM), demonstrating the stable good performance of our proposed methods.

Qualitative Comparison. Figs. 5 and 6 presents visual comparisons on the LOL dataset, MEF dataset and DICM dataset. Our model can enlighten low-light inputs correctly while successfully maintaining texture information and effectively removing noise, demonstrating the superiority of our proposed method.

Table 3. NIQE scores on the no-reference datasets, including MEF, LIME, NPE, and DICM. The best and the second results are marked in bold and underlined, respectively.

Type	Method	MEF	LIME	NPE	DICM	Ave. NIQE↓
	input	4.265	4.438	4.319	4.255	4.319
T	HE[11]	3.508	4.892	4.045	3.825	4.068
	CLAHE [26]	3.352	3.982	4.032	3.731	3.774
	LIME[8]	3.720	4.155	4.268	3.846	4.000
U	RRDNet[39]	3.480	3.911	3.975	3.972	3.835
	ZeroDCE[7]	3.467	4.130	4.233	3.652	3.871
	RUAS[20]	5.142	4.827	6.586	6.674	5.807
	SCI[23]	3.681	4.011	3.878	3.786	3.839
	PairLLE[5]	3.911	4.348	4.323	4.090	4.168
	CycleGAN[40]	3.782	**3.276**	4.036	3.560	3.664
	En.GAN[12]	3.232	3.719	4.113	3.570	3.659
	Ours	**3.215**	3.708	**3.839**	**3.513**	**3.569**
S	RetinexNet[33]	4.215	5.109	4.283	4.096	4.234

Table 4. Quantitative comparison for Parameters, FLOPs, and Runtime. Params denotes the parameter number.

Method	Params (M)	FLOPs (G)	Runtime (S)
RetinexNet[33]	0.838	148.54	0.023
MBLLEN[21]	0.450	21.37	0.159
DRBN[34]	0.577	42.41	0.140
En.GAN[12]	8.367	72.61	0.011
ZeroDCE[7]	0.0789	5.2112	0.0042
RUAS[20]	0.0014	0.2813	0.0063
SCI[23]	0.0003	0.0619	0.0017
Ours	0.6928	24.82	0.023

3.2 Ablation Studies

We perform an ablation study to verify the effectiveness of our contrastive learning strategy employed in stage II on the LOL dataset. The quantitative comparison is concluded in Table. 1, where we report the initial results of stage I, and investigate the role of FSAM, $\mathcal{L}^G_{texture}$ and $\mathcal{L}^G_{illu}$. We further provide visualizations of the enhancement results of different settings in Fig. 7. As observed, by adopting dual FSAM, contrastive term $\mathcal{L}^G_{texture}$ and $\mathcal{L}^G_{illu}$, our method significantly correct the lightness and preserves texture details, thus promoting model performance.

4 Conclusion

To tackle degradations like poor visibility, acute noise, and low contrast, we analyze the previous VGG-based low-light image enhancement methods and explore the potential of contrastive learning on refinement. Following the principle of coarse-to-fine, we disentangle the whole enhancement process into an unpaired low-light image enhancement and refinement stage. Through the utilization of the proposed Frequency-Spatial Attention Module, we acquire an adaptive illumination map to guide light enhancement and noise removal. A state-of-the-art performance has been proved on various datasets. However, our proposed method may exhibit diminished performance in extremely dark scenarios due to the inherent loss of detail and noise in low-light images. Compared to supervised methods with ground-truth supervision, our approach generates fewer structural details. Enhancements can be improved by incorporating robust semantic information. In the future, it will be worthwhile to explore training a large pre-trained model that incorporates semantic features. We hope that our strategy could serve as a new refinement scheme for future research.

Acknowledgments. This work was supported by the Anhui Provincial Natural Science Foundation under Grant 2108085UD12. We acknowledge the support of GPU cluster built by MCC Lab of Information Science and Technology Institution, USTC.

References

1. Cai, Y., Bian, H., Lin, J., Wang, H., Timofte, R., Zhang, Y.: Retinexformer: One-stage retinex-based transformer for low-light image enhancement. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 12504–12513 (2023)
2. Du, W., Chen, H., Yang, H.: Learning invariant representation for unsupervised image restoration. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14483–14492 (2020)
3. Fan, M., Wang, W., Yang, W., Liu, J.: Integrating semantic segmentation and retinex model for low-light image enhancement. In: Proceedings of the 28th ACM International Conference on Multimedia. pp. 2317–2325 (2020)
4. Fei, B., Lyu, Z., Pan, L., Zhang, J., Yang, W., Luo, T., Zhang, B., Dai, B.: Generative diffusion prior for unified image restoration and enhancement. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9935–9946 (2023)
5. Fu, Z., Yang, Y., Tu, X., Huang, Y., Ding, X., Ma, K.K.: Learning a simple low-light image enhancer from paired low-light instances. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 22252–22261 (2023)
6. Goodfellow, I.J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial networks. arXiv preprint arXiv:1406.2661 (2014)

7. Guo, C., Li, C., Guo, J., Loy, C.C., Hou, J., Kwong, S., Cong, R.: Zero-reference deep curve estimation for low-light image enhancement. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1780–1789 (2020)
8. Guo, X., Li, Y., Ling, H.: LIME: Low-light image enhancement via illumination map estimation. IEEE Trans. Image Process. **26**(2), 982–993 (2016)
9. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: Gans trained by a two time-scale update rule converge to a local nash equilibrium. Adv. Neural. Inf. Process. Syst. **30**, 1–12 (2017)
10. Huang, J., Liu, Y., Zhao, F., Yan, K., Zhang, J., Huang, Y., Zhou, M., Xiong, Z.: Deep fourier-based exposure correction network with spatial-frequency interaction. In: Proceedings of the 17th European Conference on Computer Vision. pp. 163–180 (2022a)
11. Ibrahim, H., Kong, N.S.P.: Brightness preserving dynamic histogram equalization for image contrast enhancement. IEEE Trans. Consum. Electron. **53**(4), 1752–1758 (2007)
12. Jiang, Y., Gong, X., Liu, D., Cheng, Y., Fang, C., Shen, X., Yang, J., Zhou, P., Wang, Z.: Enlightengan: Deep light enhancement without paired supervision. arXiv preprint arXiv:1906.06972 (2019)
13. Ke, R., Schönlieb, C.B.: Unsupervised image restoration using partially linear denoisers. IEEE Trans. Pattern Anal. Mach. Intell. **44**(9), 5796–5812 (2021)
14. Land, E.H.: The retinex theory of color vision. Sci. Am. **237**(6), 108–129 (1977)
15. Lee, C., Lee, C., Kim, C.S.: Contrast enhancement based on layered difference representation. In: Proceedings of the International Conference on Image Processing. pp. 965–968 (2012)
16. Li, C., Guo, C.L., Zhou, M., Liang, Z., Zhou, S., Feng, R., Loy, C.C.: Embedding fourier for ultra-high-definition low-light image enhancement. arXiv preprint arXiv:2302.11831 (2023)
17. Li, L., Wang, R., Wang, W., Gao, W.: A low-light image enhancement method for both denoising and contrast enlarging. In: Proceedings of the IEEE International Conference on Image Processing. pp. 3730–3734 (2015)
18. Lin, X., Ren, C., Liu, X., Huang, J., Lei, Y.: Unsupervised image denoising in real-world scenarios via self-collaboration parallel generative adversarial branches. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 12642–12652 (2023)
19. Lin, X., Yue, J., Ding, S., Ren, C., Guo, C.L., Li, C.: Unlocking low-light-rainy image restoration by pairwise degradation feature vector guidance. arXiv preprint arXiv:2305.03997 (2023)
20. Liu, R., Ma, L., Zhang, J., Fan, X., Luo, Z.: Retinex-inspired unrolling with cooperative prior architecture search for low-light image enhancement. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10561–10570 (2021)
21. Lv, F., Lu, F., Wu, J., Lim, C.: MBLLEN: Low-light image/video enhancement using CNNs. In: Proceedings of the British Machine Vision Conference. pp. 1–13 (2018)
22. Ma, K., Zeng, K., Wang, Z.: Perceptual quality assessment for multi-exposure image fusion. IEEE Trans. Image Process. **24**(11), 3345–3356 (2015)
23. Ma, L., Ma, T., Liu, R., Fan, X., Luo, Z.: Toward fast, flexible, and robust low-light image enhancement. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5637–5646 (2022)

24. Mittal, A., Soundararajan, R., Bovik, A.C.: Making a "completely blind" image quality analyzer. IEEE Signal Process. Lett. **20**(3), 209–212 (2012)
25. Park, S., Yu, S., Kim, M., Park, K., Paik, J.: Dual autoencoder network for retinex-based low-light image enhancement. IEEE Access **6**, 22084–22093 (2018)
26. Pizer, S.M., Amburn, E.P., Austin, J.D., Cromartie, R., Geselowitz, A., Greer, T., ter Haar Romeny, B., Zimmerman, J.B., Zuiderveld, K.: Adaptive histogram equalization and its variations. Computer vision, graphics, and image processing **39**(3), 355–368 (1987)
27. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)
28. Triantafyllidou, D., Moran, S., McDonagh, S., Parisot, S., Slabaugh, G.: Low light video enhancement using synthetic data produced with an intermediate domain mapping. In: Proceedings of the European Conference on Computer Vision. pp. 103–119. Springer (2020)
29. Tsin, Y., Ramesh, V., Kanade, T.: Statistical calibration of ccd imaging process. In: Proceedings of IEEE International Conference on Computer Vision. vol. 1, pp. 480–487 (2001)
30. Wang, S., Zheng, J., Hu, H.M., Li, B.: Naturalness preserved enhancement algorithm for non-uniform illumination images. IEEE Transactions on Image Processingg **22**(9), 3538–3548 (2013)
31. Wang, X., Yu, K., Wu, S., Gu, J., Liu, Y., Dong, C., Qiao, Y., Change Loy, C.: ESR-GAN: Enhanced super-resolution generative adversarial networks. In: Proceedings of the European Conference on Computer Vision Workshops. pp. 63–79 (2018)
32. Wang, Y., Cao, Y., Zha, Z.J., Zhang, J., Xiong, Z., Zhang, W., Wu, F.: Progressive retinex: Mutually reinforced illumination-noise perception network for low-light image enhancement. In: Proceedings of the 27th ACM International Conference on Multimedia. pp. 2015–2023 (2019)
33. Wei, C., Wang, W., Yang, W., Liu, J.: Deep retinex decomposition for low-light enhancement. arXiv preprint arXiv:1808.04560 (2018)
34. Yang, W., Wang, S., Fang, Y., Wang, Y., Liu, J.: From fidelity to perceptual quality: A semi-supervised approach for low-light image enhancement. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3063–3072 (2020)
35. Zhang, L., Liu, X., Learned-Miller, E., Guan, H.: Sid-nism: A self-supervised low-light image enhancement framework. arXiv preprint arXiv:2012.08707 (2020)
36. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 586–595 (2018)
37. Zhang, Y., Zhang, J., Guo, X.: Kindling the darkness: A practical low-light image enhancer. In: Proceedings of the 27th ACM International Conference on Multimedia. pp. 1632–1640 (2019)
38. Zhang, Y., Di, X., Zhang, B., Wang, C.: Self-supervised image enhancement network: Training with low light images only. arXiv preprint arXiv:2002.11300 (2020)
39. Zhu, A., Zhang, L., Shen, Y., Ma, Y., Zhao, S., Zhou, Y.: Zero-shot restoration of underexposed images via robust retinex decomposition. In: Proceedings of the IEEE International Conference on Multimedia and Expo. pp. 1–6 (2020)
40. Zhu, J.Y., Park, T., Isola, P., Efros, A.A.: Unpaired image-to-image translation using cycle-consistent adversarial networks. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 2223–2232 (2017)

Transformer-Based Fringe Restoration for Shadow Mitigation in Fringe Projection Profilometry

Vaishnavi Ravi(✉), Siddharth Parlapalli, Sameer Ranjan, and Rama Krishna Gorthi

Indian Institute of Technology, Tirupati, Tirupati, Andhra Pradesh, India
vaishnavi1712@gmail.com

Abstract. Fringe Projection Profilometry (FPP) is a widely recognized technique for deriving 3D profiles from images. Despite numerous methodologies developed to determine depth in FPP, the inherent triangulation setup of camera, projector, and object often introduces substantial shadows in captured fringes, especially for complex objects. These shadows can impede algorithm performance and introduce undesirable artifacts in final depth profiles. In this work, we introduce a Transformer-based Fringe Restoration network designed to repair shadowed regions in single deformed fringe images. The network comprises an object localization module to identify object regions and a shadow repair module that utilizes reference and deformed fringes to restore shadowed areas. In addition, we construct a comprehensive pseudo-realistic dataset using Blender, a computer graphics tool, to train the proposed network. Our results demonstrate precise object region segmentation with just a single fringe image, and the proposed network achieves superior fringe restoration, as quantified by Intersection over Union (IoU) and dice score metrics. Moreover, 3D reconstruction on shadowed and shadow-free deformed fringes using standard single-shot methods exhibits enhanced performance owing to the fringe restoration network.

Keywords: Shadow Removal · Depth Estimation · Fringe Projection Profilometry

1 Introduction

Fringe Projection Profilometry (FPP) stands out as a robust method within optical metrology, enabling precise and non-contact 3D surface measurements with remarkable accuracy and resolution. Its applications span across diverse industries such as automotive, aerospace, biomedical, cultural heritage preservation etc. [1]

A typical FPP system comprises a projector that shines a sinusoidal pattern of light onto the object under examination and a camera that captures the

S. Parlapalli and S. Ranjan—These authors contributed equally to this work.

A. Antonacopoulos et al. (Eds.): ICPR 2024, LNCS 15321, pp. 339–354, 2025.
https://doi.org/10.1007/978-3-031-78305-0_22

deformations of this pattern from a different view. The prevalent configuration involves a cross-axis symmetric arrangement, wherein either the projector, the camera, or both are tilted to align their fields of view. However, due to this arrangement and the presence of a reference plane behind the object, shadows occur in the captured deformed fringes.

These shadows pose a significant challenge, as they adversely affect the accuracy and reconstruction capability of the 3D profiling process. They result in missing data, erroneous depth estimates, and can also introduce artifacts. Consequently, researchers have proposed various strategies to mitigate the shadow-related issues in FPP systems. Many of these approaches rely on image processing techniques although their effectiveness may be limited when applied to fringe images.

1.1 Motivation and Contribution

Nowadays, deep learning (DL) based models have demonstrated state-of-the-art performance over existing conventional algorithms in various low-vision tasks such as image enhancement, restoration, object detection, and 3D reconstruction.

The main motivation of this work is to employ a learning-based framework which repairs the shadows present in deformed fringes using a single-shot approach, thereby estimating precise 3D profiles in FPP. An overall high-level block diagram is given in the Fig.1 below.

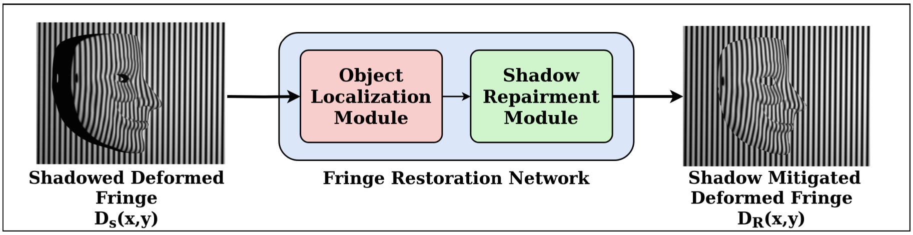

Object Localization Module
Shadow Repairment Module
Shadowed Deformed Fringe
$D_S(x,y)$
Fringe Restoration Network
Shadow Mitigated Deformed Fringe
$D_R(x,y)$

Fig. 1. Overall Block Diagram for Proposed Fringe Restoration Network

The main contributions of this work are:

- Proposed a Transformer-based Fringe Restoration network that takes a single deformed fringe to repair the shadow regions.
- The proposed Fringe Restoration network encompasses an object localization module to estimate the object mask and shadow repairment module to mitigate the shadow in deformed fringes.
- A comprehensive dataset of fringe images with realistic shadow effects is created using a computer graphics tool Blender for effective training of the Fringe Restoration network.
- Proposed object localization module is evaluated in terms of Mean Absolute Error (MAE), Dice Score and IoU score on synthetic and real test samples.

– The proposed method's 3D profiling results on shadowed images are evaluated using well-known single-shot approaches like Fourier Transform Profilometry (FTP) and Windowed FTP (WFTP) followed by Quality Guided Phase Unwrapping (QGPU).

The remainder of the paper is structured as follows. Sec.2 reviews the existing literature on shadow-repairing techniques. Sec.3 describes the problem formulation, proposed approach, data generation procedure and loss functions. Sec.4 presents the implementation details and results. Following this, the conclusions are presented in Sec.5.

2 Related Literature

Numerous efforts have been made by researchers to address the challenge of shadow elimination in earlier studies. Initially, some explored the incorporation of multiple projectors or cameras from different perspectives to combine information from multiple sources to deal with shadows [2], [3], [4]. These methods were useful but involved high costs due to increased equipment.

Subsequently, the focus shifted towards developing image processing-based solutions to separate the shadow from the background. While conventional methods for natural images relied on evaluating brightness and chrominance information [5], distinguishing shadows in fringe images, characterized by specific patterns and intensities akin to black fringes, proved challenging. Zhang [6] proposed hole detection and filling method that employed a Gaussian filter to smooth fringe patterns. However, this approach inadvertently smoothed object surfaces, blurring reconstructed model details. Chen et al. [7] employed two-dimensional linear interpolation to derive a final phase map after eliminating invalid points' phases. Yet, indiscriminate smoothing of shadow areas introduced additional phase errors across the repaired surface. Huang et al. [8] developed an identification framework to remove the phase of invalid regions in the image like shadows. Later, Lu, Zhang, and Zhong et al. [9], [10], [11] proposed an approach to delete the phase in shadow area using an intensity-based threshold on the modulation images. However, such methods often lacked precision.

Following this, Otsu's thresholding is applied to solve the shadow problem in [12]. Wang et al. proposed a method using a k-means clustering approach to segment the background from the shadow regions. But, these two methods are very time-consuming [13]. In [14], Zheng et al. utilized intensity difference between the phase of shifting patterns to distinguish between shadow and background. Lu et al. [9] proposed to map a 3D point cloud onto the digital micromirror device (DMD) plane and then identify the shadow regions by comparing the captured fringe images with the mapping results. Precise system calibration is required for this method. Another approach combined image gradient squares, binary image sub-region areas, and image decomposition/composition to eliminate invalid phase values. But then these also involve intense computations. After this, Lopez et al. proposed a method in [15] where the insufficient data in shaded areas are substituted by masked reference planes and this segmentation is achieved by superpixel-based fast fuzzy c-means clustering algorithm. This also involves intense calculations. In another attempt to classify the shadow areas into valid and invalid, in which the valid shadow area is repaired by a neighboring information fusion phase estimation (NIFPE), and the invalid area is repaired by background phase matching (BPM) algorithms, respectively.

Recently DL methods have shown potential results to deal with shadows in natural images as given in [16] and [17]. Inspired by this, some researchers have developed DL-based methods for shadow removal in deformed fringe images. In [18], Wang et al. proposed a direction-aware spatial context module based network coupled with a generative adversarial network (GAN) for shadow region detection and repair. It identifies the shadow regions in the image. Following this, Li et al. introduced TPDNet, a DL model trained to estimate depth maps from texture images, masks, and unwrapped phase maps. However, acquiring object masks poses challenges in real-world images.

Hence, we propose a Transformer-based Fringe Restoration network that takes a single deformed fringe to repair the shadow regions. The proposed network has an object localization module which identifies the object region in the images, and a shadow repairment module which restores the shadow regions using reference and deformed fringes.

3 Proposed Methodology

In this section, we provide a comprehensive overview of problem formulation where the importance of shadow repairment is discussed, followed by the description of the proposed DL-based Fringe Restoration network for repairing the shadows. After this, the complete data generation procedure with real shadow effects using Blender for training the proposed DL-based model and implementation details are elaborated.

3.1 Problem Formulation

In FPP, a computer-generated fringe pattern is projected on the object of interest, and the deformations are captured by a camera from another view as given in Fig.2. The equation for fringe pattern can be written as:

$$I_d(x,y) = a(x,y) + b(x,y)cos(2\pi f_c x + \Phi(x,y)) \tag{1}$$

where a is the average intensity variations in the background, b is the non-uniform reflectivity of the diffusely reflecting object, $\Phi(x,y)$ is the phase term introduced by the object's height profile, which is proportional to the shape of the object. There are many algorithms developed to extract the height information.

Conventional FPP methods involve various steps like fringe analysis - to extract the wrapped phase and phase unwrapping - to retrieve the absolute phase from the wrapped ones, followed by calibration which maps the absolute phase to height as given in [1]. Most works in the literature use Phase Shifting

Profilometry (PSP) [19] followed by Multi-Frequency Temporal Phase Unwrapping (MFTPU) [20] which takes images of multiple frequencies and phase shifts to extract the absolute phase shift. On the otherhand, FTP [21] followed by QGPU is used in place of PSP + MFTPU to address the dynamic scenarios using a single fringe. Finally, these absolute phase shifts are converted to precise height information with the proportionality constant given by the approaches in [22].

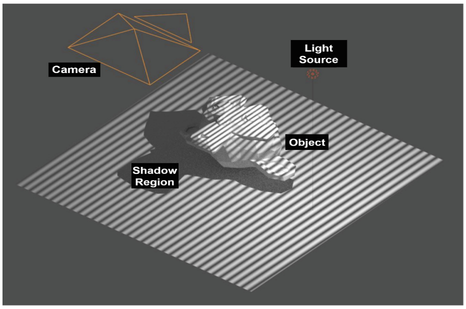


Fig. 2. Problem Formulation

But, because of the triangulation principle, objects cater shadows onto the reference plane present in the scene. These shadows recorded in captured fringes hinder the performance of reconstruction algorithms and can introduce artifacts in the wrapped phase maps. If a path-following algorithm like QGPU is employed for unwrapping, it completely fails because of the presence of shadows. So, there is a primary requirement to repair the shadows in the captured images using a single-shot approach. Therefore, we posed this fringe restoration task as a segmentation problem to localize the object region and developed a Transformer-based DL model followed by a shadow repairment to restore the shadow regions. In this work, we also generated a rich dataset of fringe images with realistic effects like shadows to train the DL model whose details are given in Sec.3.3.

3.2 Proposed Transformer-based Fringe Restoration Network

In this subsection, we present the details of the proposed Fringe Restoration network for shadow removal. Recently, a GAN-based approach for shadow repairment focused on identifying the shadow region in the fringe image and repairing it is proposed. But, our approach is different from [18] in the way that it has two modules, namely (1) Object Localization module - which detects the object region first unlike the shadow mask in [18] and then (2) Shadow Repairment module which repairs the captured fringe leveraging the already available reference and deformed fringes.

Object Localization module: The block diagram of the Object Localization module is shown in Fig.3. Let the shadowed deformed fringe image be represented as $D_s(x, y)$ and the object mask is $M(x, y)$. This module's task is to localize the object region from a single $D_s(x, y)$ and generate $M(x, y)$. This object mask generation task is posed as a learning-based segmentation problem. We have leveraged a recent transformer-based segmentation framework, Swin Transformer [23] for performing this task. The key design element of this transformer is its shift of window partition between the self-attention layers. These bridge the windows of the preceding layer, providing connections among them which enhances the modeling power. This feature is expected to be instrumental in accumulating the object pixels and segmenting out the object.

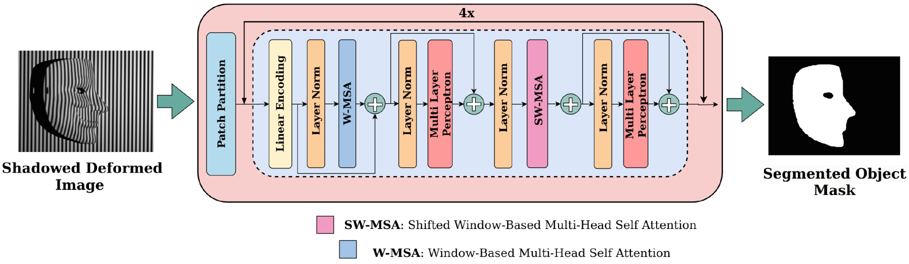


Fig. 3. Block Diagram of Object Localization Module

The architecture of Swin Transformer consists of patch partition, which divides the image into small regions and feeds to 4 concatenated stages of linear embedding and Swin Transformer block connected back-to-back. The Swin Transformer block consists of a shifted windows-based multi-head self-attention (MSA) followed by a multilayer perceptron (MLP) with Gaussian Error Linear Unit Activation (GELU) non-linearity in between. A LayerNorm is applied before each MSA module and each MLP, and also a residual connection is applied after each module. The segmented object mask obtained at the end of this module is fed to the Shadow Repairment Module.

Shadow Repairment module: The segmented object mask $M(x, y)$ predicted by the first module acts as the input to this module. Along with the mask, this module takes the reference and the shadowed deformed fringe and performs the shadow repairment as given in Fig.4. If the reference and the deformed fringes are considered as $R(x, y)$ and $D_S(x, y)$ respectively then the restored fringe image $D_R(x, y)$ can be obtained by the equation given below:

$$D_R(x, y) = [\sim M(x, y) * R(x, y)] + [M(x, y) * D_S(x, y)] \tag{2}$$

where $[\sim M(x, y)]$ is the negative of object mask $M(x, y)$. Then, the final $D_R(x, y)$ obtained at the end of this module can be observed to be free from shadow regions retrieved from just one single deformed fringe. However, the basic assumption on which this module works is that the intensity variations of the reference $R(x, y)$ and the shadowed deformed fringes are the same and there are no illumination variations while recording the images. This ensures that the shadow-mitigated deformed fringe will have uniform intensities all over. The block diagram of this module is given in Fig.4.

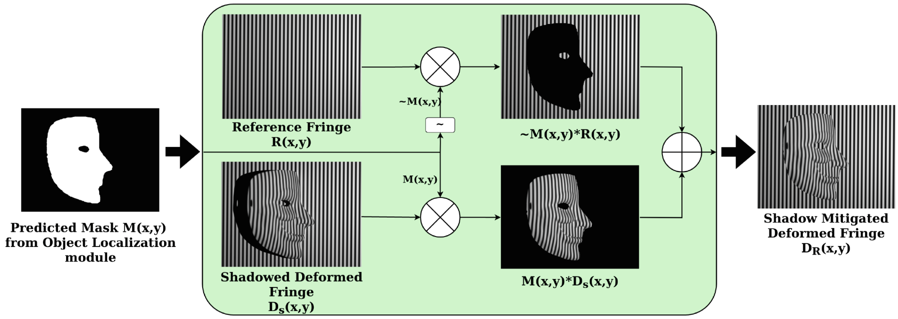


Fig. 4. Block Diagram of Shadow Repairment

Loss Function The loss function employed to train the Swin Transformer-based segmentation network mentioned in the first module is Binary Cross Entropy (BCE) calculated using the following expression:

$$L(y, \hat{y}) = -\frac{1}{N}\sum_{i=1}^{N}[y_i \log(\hat{y}_i) + (1 - y_i)\log(1 - \hat{y}_i)] \tag{3}$$

where $L(y, \hat{y})$ is the BCE loss in which y_i is the true label and $\hat{y}_i$ is the predicted probability of the i^{th} sample belonging to class 1, respectively. The total number of samples is given by N.

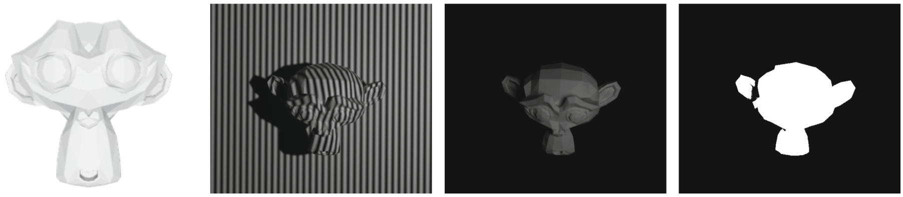
3.3 Pseudo-Realistic Data Generation

It is well-known that the DL networks are data-hungry and need large amounts of data to train for their effective learning.

As collecting diverse real training samples is a very time-consuming and expensive process, it is preferred to generate synthetic datasets as performed in earlier works [24], [25], [26]. However, the task at hand is shadow removal and hence data cannot be generated using the procedures mentioned in these works. Further, it is not straightforward to create shadows through mathematical modeling. Hence, we have adopted a pseudo-realistic data generation procedure with the help of a Computer Graphics tool called "Blender". Many recent works have employed this kind of data-generation procedure for various tasks [18], [27]. In this work, we employed Blender to generate shadowed deformed fringes for the

3D Computer-Aided Design (CAD) objects and the object mask. This dataset can be accessed from (*Dataset*) and can be employed to train the Object Localization module.

A virtual FPP system with a light source, camera and a reference plane as a wall is arranged in Blender which imitates the original FPP system. The 3D models of the objects are sourced from various prominent repositories and datasets available online. One such sample is shown in Fig. 5. To make the captured data realistic, various parameters like the pitch, noise, material properties, type of light source etc. have been carefully set.



(a) CAD Model (b) Shadowed Deformed Fringe (c) Without Reference Plane (d) Ground Truth

Fig. 5. Training Data Generation in Blender.

The step-by-step procedure to set-up FPP virtually in Blender is as follows:

- **Camera** & **Projector**: The camera in the scene is set to the "Perspective projection" mode. The light source chosen to be "Spotlight" is positioned and oriented to project fringes onto a reference plane. The projector is tilted at an angle that mimics the intensity falloff observed in the real-world camera setup. The camera and projector are along the same line to get the shadowing effect. The “spotlight" is replaced with a "point source" and the reference plane is removed to obtain image as shown in Fig.5(c) which when thresholded as shown in Fig.5(d) gives a ground truth mask to train the Object Localization module.
- **Data Samples**: Each 3D object is scaled to fit in camera's FOV. In order to diversify the dataset, each object is set to 10 rotations along X,Y and Z axes along with projecting patterns with 4 phase shifts of 0°, 45°, 90° and 135°. This results in 120 samples from a single 3D CAD object.
- **Automated Pipeline**: The Blender Python package (bpy) played a crucial role in automating the entire data generation pipeline which generated the deformed, reference, and object masks in this work.

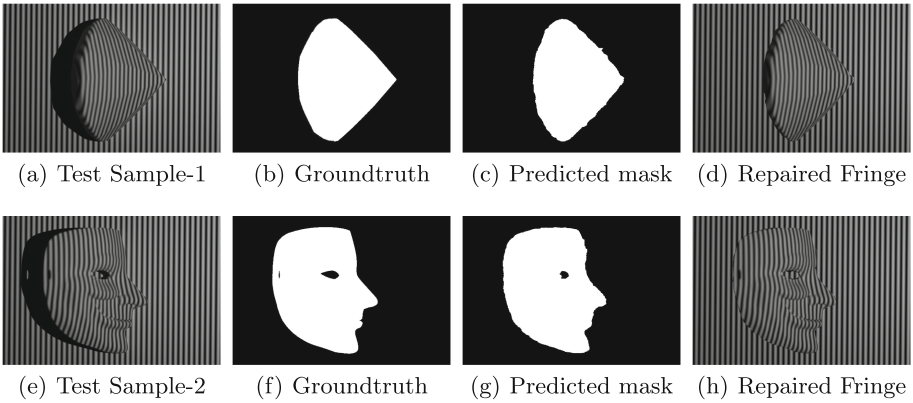
To ensure a robust training process, the entire dataset is partitioned into mutually exclusive train, validation, and test sets in 8:1:1 ratio. This strategic mix aims to equip the model with the versatility needed to accurately reconstruct real-world objects with diverse profiles.

3.4 Implementation Details

A learning rate of 1×10^{-3} with multi-step learning rate scheduler and Adam optimizer with $\beta_1 = 0.9$ and $\beta_2 = 0.999$ are used. A total of 45 3D objects have been selected from web sources and 5400 shadowed deformed fringe images and their corresponding object masks are generated. The model was trained on 4320 training and 540 validation samples for 30 epochs. The training was performed on 4 NVIDIA A100-SXM Graphical Processing Unit (GPU). The results and evaluations presented in this paper are tested with final weights obtained after the entire process.

4 Results

In this section, we present the results of the proposed Fringe Restoration Network on various synthetic and real samples. The results of the Object Localization module is evaluated using metrics like Mean Absolute Error (MAE), IoU score and Dice Score. Following this, in order to prove the efficacy of the proposed method, the shadowed and repaired deformed fringes are reconstructed using two standard single shot fringe reconstruction methods like Fourier Transform Profilometry (FTP) [21] and Windowed FTP (WFTP) [28] followed by QGPU [29].


(a) Test Sample-1 (b) Groundtruth (c) Predicted mask (d) Repaired Fringe

(e) Test Sample-2 (f) Groundtruth (g) Predicted mask (h) Repaired Fringe

Fig. 6. Test images, ground truth, and predictions from the proposed method.

4.1 Results of the Object Localization Module

This subsection presents the results of the Object Localization and Shadow repairment modules for synthetic test samples.

With the proposed fringe restoration network, the object region is localized and the shadows are repaired as shown in Fig.6. Two synthetic test samples with significant shadow regions are shown in Fig.6(a) and Fig.6(e). The ground truth object masks for these test samples are given in Fig.6(b) and Fig.6(f). The predicted results of the first module i.e, object localization module is in Fig.6(c) and Fig.6(g). It can be observed that the Swin Transformer model is trained very well to segment out the object regions very precisely. Then, the final results of the shadow repairment module are presented in Fig.6(d) and Fig.6(h). It can be observed that the shadows are precisely removed and intensities in the shadow regions are properly restored. In order to quantify the performance of the proposed module, we have calculated 3 metrics namely MAE, IoU, and Dice score. These are calculated between the predicted mask, and the ground truth and are presented in Table.1. The values reported in Table.1 show that the model has been trained well to identify the object location precisely.

Table 1. Evaluation Metrics for Mask

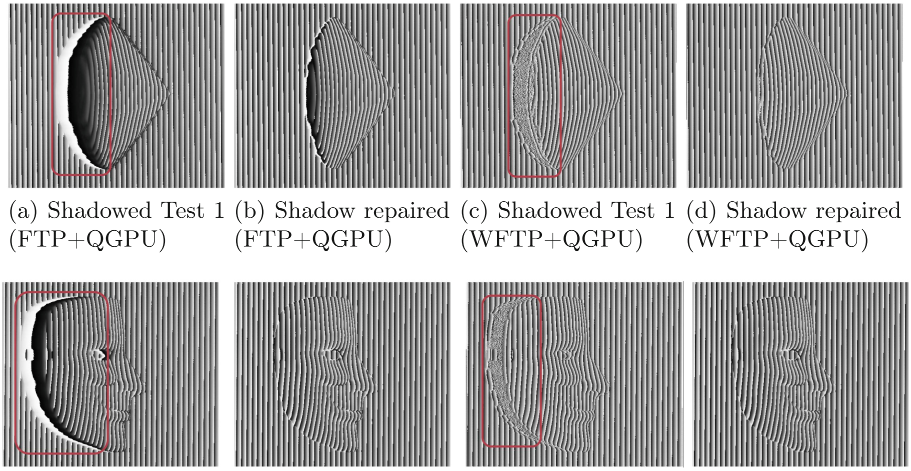
Metric	**Mean Absolute Error (MAE)**	**Intersection over Union (IoU Score)**	**Dice Score**
Proposed Method	0.002	0.949	0.974

4.2 Evaluation Results on Synthetic Samples for 3D Profiling

Evaluation in terms of wrapped phases: Here, we present the wrapped phase results of two test samples shown in Fig.6(a) and Fig.6(d). Fig.7 represents the wrapped phase maps obtained using FTP and WFTP on shadowed and shadow-repaired fringe images. It can be seen that the shadowed deformed fringes have lots of artifacts in the wrapped phase maps, especially in the highlighted regions. These artifacts are highlighted in Fig.7(a), 7(e), 7(c), 7(g) and act as potential sources of error propagation while performing Phase Unwrapping using single-shot unwrapping methods like QGPU.

Evaluation of test samples in terms of 3D profiles: In this subsection, the absolute phase maps of the test samples are generated using QGPU on wrapped phase results of the test objects shown in Fig.7.

The actual 3D profiles, which are proportional to these absolute phase maps, are generated for shadowed and shadow-repaired fringes to clearly depict the effectiveness of the proposed Fringe Restoration network. This can be validated by the MAE values given in Table.2 where the 3D profiles of samples in the test dataset are generated and the MAEs are calculated in shadowed and shadow-repaired cases. From the table, we can observe that there is a 43% and 57% reduction in the error values of FTP+QGPU and WFTP+QGPU, respectively, which validates the fact that repairing shadows indeed helps in the improvement of the final 3D profile.



(a) Shadowed Test 1 (FTP+QGPU) (b) Shadow repaired (FTP+QGPU) (c) Shadowed Test 1 (WFTP+QGPU) (d) Shadow repaired (WFTP+QGPU)

(e) Shadowed Test 2 (FTP+QGPU) (f) Shadow repaired (FTP+QGPU) (g) Shadowed Test 2 (WFTP+QGPU) (h) Shadow repaired (WFTP+QGPU)

Fig. 7. Wrapped phases of shadowed, shadow-free fringes with FTP and WFTP.

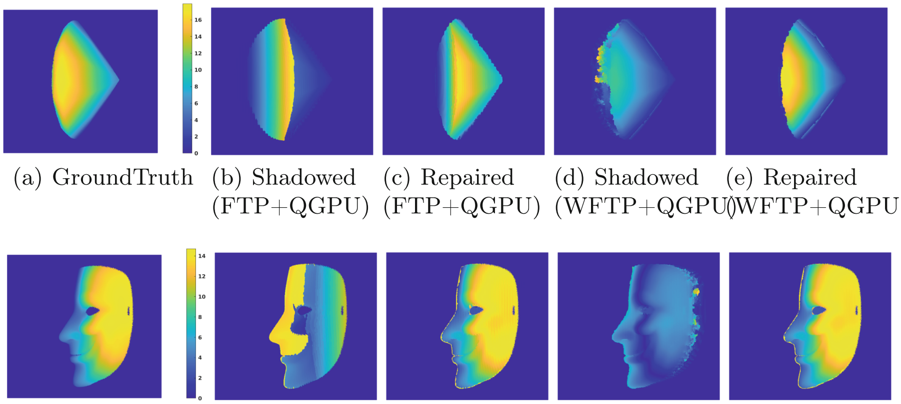
Table 2. MAE values for Shadowed and Shadow-Repaired images absolute phase values for test data samples

Method	**Shadowed (MAE)**	**Shadow-Free (MAE)**
FTP+QGPU	6.136	3.503
WFTP+QGPU	6.337	2.709

To demonstrate the performance visually, the 3D profiles of two test samples are shown in Fig.8. The ground truth surface profiles of the 3D test objects are given in Fig.8(a), Fig.8(b). It can be clearly observed that the edges on the object near the shadow regions incur high errors, and there are a few places in the shadowed test samples where the unwrapping path is missed, leading to unusual errors in the absolute phase maps. Once the shadows are repaired, the absolute phase maps become more accurate as QGPU performs well.

4.3 Experimental Comparison with the State-of-the-Art

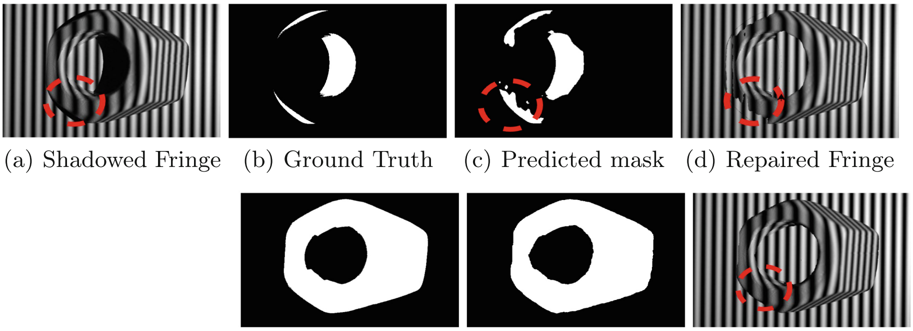
One unique distinction of the proposed approach is that *we estimate object mask in the process of dealing with shadows in FPP, whereas the existing methods like* [18] *estimate a shadow mask.* Further, for the shadow detection and removal in [18], a mask is estimated for the shadow region using a complex multistep process. This could not be automated in an end-to-end fashion as it has two distinct DL models to train and also several handcrafted steps between the two models. *In contrast, the proposed method is simple and straightforward, which*



(a) GroundTruth (b) Shadowed (FTP+QGPU) (c) Repaired (FTP+QGPU) (d) Shadowed (WFTP+QGPU)(e) Repaired (WFTP+QGPU)

(f) GroundTruth (g) Shadowed (FTP+QGPU) (h) Repaired (FTP+QGPU) (i) Shadowed (WFTP+QGPU)(j) Repaired (WFTP+QGPU)

Fig. 8. Groundtruths, (FTP/WFTP + QGPU) estimated 3D profiles of Gem and Facemask samples with shadow and shadow repaired results.



(a) Shadowed Fringe (b) Ground Truth (c) Predicted mask (d) Repaired Fringe

(e) Ground Truth (f) Predicted Mask (g) Repaired Fringe

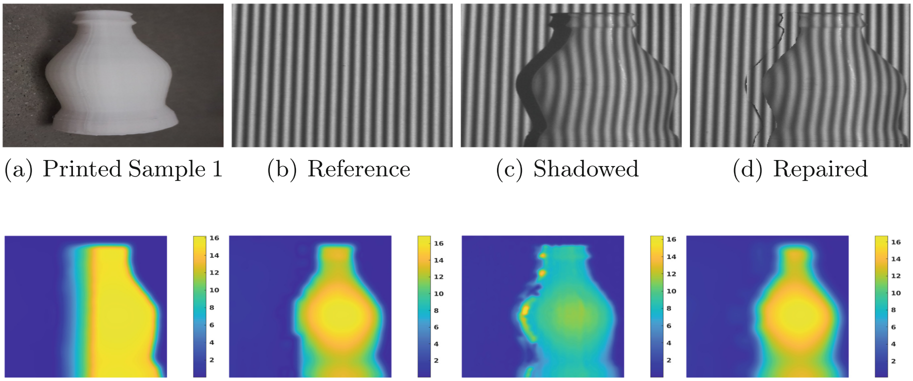
Fig. 9. Comparison of shadow mask and object mask-based approaches. The first row shows the results of the shadow mask-trained model as in existing frameworks [18]. The second row shows the results of the object mask-trained model for a test sample, as in the proposed framework.

performs shadow removal with a single end-to-end trained transformer network. In this subsection, we compare the idea of shadow mask as employed in [18] versus the object mask prediction strategy proposed in this work using the same transformer-based model and discuss the pros and cons below. we generated shadow mask ground truth for our dataset by using a point source in a blender followed by thresholding. These shadow masks are used to train the same transformer model used in the Object Localization module for this experimental study.

In highlighted regions, small shadows adjacent to fringe troughs caused the shadow mask-trained model to fail, producing faulty masks (Fig. 9(b)). In contrast, our Object Localization module estimated better object masks, resulting in higher quality repaired fringe (Fig. 9(g)). This distinction issue can also affect performance on images without shadow-deformed fringes. Small errors in shadow masks can lead to errors in the object region and depth prediction. Our approach avoids these issues unless the object mask covers a significant shadow region. Therefore, we believe object masking is superior to shadow masking.

4.4 Evaluations on Real Samples

In this subsection, we present the results on two real samples that are captured using an FPP setup in a lab environment. The experimental setup for capturing real fringes has a DLP (model: DLP Lightcrafter4500) and an industrial camera (model:DFK 27BUJ003). The projector and camera are arranged in a cross-axis symmetric arrangement with a baseline of 0.3m, a distance of baseline to the reference plane of 0.9m, and a fringe frequency of 40 cycles/row.

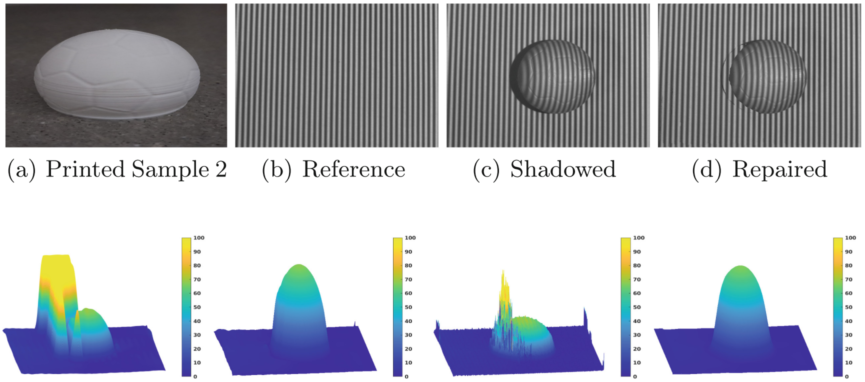


(a) Printed Sample 1 (b) Reference (c) Shadowed (d) Repaired

(e) FTP+QGPU (Shadowed) (f) FTP+QGPU (Repaired) (g) WFTP+QGPU (Shadowed) (h) WFTP+QGPU (Repaired)

Fig. 10. Results and comparisons on Real test sample 1.

Two CAD samples from our test set are fabricated into real objects through 3D printing for testing of real cases as depicted in Fig.10(a), Fig.11(a). For the first sample, the real reference and the shadowed deformed fringes are shown in Fig.10(b) and Fig.10(c). To match the contrast of real images with that of the training dataset, gamma correction is performed as a preprocessing step. Fig.10(d) shows the repaired deformed fringe estimated by the proposed Fringe Restoration network. Following this, reconstruction algorithms FTP+QGPU and WFTP+QGPU are applied on the shadowed and repaired real samples and the results are presented in Fig.10(e)-10(h). From the reconstruction results of FTP+QGPU, it can be observed in Fig.10(e) that the reconstruction of the bottle around the shadow has completely failed because of the absence of wraps in that region leading to a faulty depth profile whereas the 3D profile obtained from the repaired fringes as in Fig.10(f) reconstructs the shape more accurately. In the case of WFTP+QGPU, there are erroneous values in the depth profile obtained from the shadowed fringe as given in Fig.10(g); this is due to the addition of artifacts while applying Fourier transform locally in WFTP and subsequently leading to erroneous path selection by QGPU during unwrapping. On the other hand, better reconstruction results are obtained in the case of repaired fringes. The same analysis is carried out for the second test sample as shown in Fig.11 with a higher dynamic range (50mm), nearly double the height of sample 1. It can be clearly observed that the shadow mitigation by the proposed method is aiding in accurate depth prediction.



(a) Printed Sample 2 (b) Reference (c) Shadowed (d) Repaired

(e) FTP+QGPU (Shadowed) (f) FTP+QGPU (Repaired) (g) WFTP+QGPU (Shadowed) (h) WFTP+QGPU (Repaired)

Fig. 11. Results and comparisons on real test sample 2.

5 Conclusions

In this work, we proposed a Transformer-based Fringe Restoration network designed to mitigate the impact of shadows on depth estimation. The proposed model consists of an object localization module and a shadow repairment module which identifies the object region from shadowed deformed fringes and restores shadowed areas using reference and deformed fringes. In addition, we created a rich pseudo-realistic dataset of fringe images with realistic shadow effects using Blender for comprehensive model training. Through rigorous evaluation, our proposed model showcases precise object segmentation in fringe restoration, as quantified by metrics like MAE, IoU and dice score. Moreover, the enhanced performance of the proposed method is evidenced by improved 3D reconstruction results obtained from shadowed to shadow-free deformed fringes with standard single-shot methods. Overall, this work presents a promising approach to enhancing the accuracy and robustness of depth estimation in presence of shadows in FPP, with potential applications in various fields requiring precise 3D surface measurements.

References

1. Gorthi, S.S., Rastogi, P.: Fringe projection techniques: whither we are? Opt. Lasers Eng. **48**(2), 133–140 (2010)
2. Skydan, O.A., Lalor, M.J., Burton, D.R.: Using coloured structured light in 3-d surface measurement. Opt. Lasers Eng. **43**(7), 801–814 (2005)
3. T. Tao, Q. Chen, J. Da, S. Feng, Y. Hu, and C. Zuo, "Real-time 3-d shape measurement with composite phase-shifting fringes and multi-view system," *Optics express*, vol. 24, no. 18, pp. 20 253–20 269, 2016
4. Servin, M., Padilla, M., Garnica, G., Gonzalez, A.: Profilometry of three-dimensional discontinuous solids by combining two-steps temporal phase unwrapping, co-phased profilometry and phase-shifting interferometry. Opt. Lasers Eng. **87**, 75–82 (2016)
5. V. Ramesh *et al.*, "A class of photometric invariants: Separating material from shape and illumination," in *Proceedings Ninth IEEE International Conference on Computer Vision*. IEEE, 2003, pp. 1387–1394
6. S. Zhang, "Phase unwrapping error reduction framework for a multiple-wavelength phase-shifting algorithm," *Optical Engineering*, vol. 48, no. 10, pp. 105 601–105 601, 2009
7. F. Chen, X. Su, and L. Xiang, "Analysis and identification of phase error in phase measuring profilometry," *Optics express*, vol. 18, no. 11, pp. 11 300–11 307, 2010
8. Huang, L., Asundi, A.K.: Phase invalidity identification framework with the temporal phase unwrapping method. Meas. Sci. Technol. **22**(3), 035304 (2011)
9. Lu, L., Xi, J., Yu, Y., Guo, Q., Yin, Y., Song, L.: Shadow removal method for phase-shifting profilometry. Appl. Opt. **54**(19), 6059–6064 (2015)
10. Zhang, S.: Composite phase-shifting algorithm for absolute phase measurement. Opt. Lasers Eng. **50**(11), 1538–1541 (2012)
11. Zhong, K., Li, Z., Shi, Y., Wang, C., Lei, Y.: Fast phase measurement profilometry for arbitrary shape objects without phase unwrapping. Opt. Lasers Eng. **51**(11), 1213–1222 (2013)
12. Xiao, Y., Li, Y.-F.: Improved 3d measurement with a novel preprocessing method in dfp. Robotics and Biomimetics **4**, 1–11 (2017)
13. Wang, H., Kemao, Q., Soon, S.H.: Valid point detection in fringe projection profilometry. Opt. Express **23**(6), 7535–7549 (2015)
14. Zheng, D., Da, F., Kemao, Q., Seah, H.S.: Phase error analysis and compensation for phase shifting profilometry with projector defocusing. Appl. Opt. **55**(21), 5721–5728 (2016)
15. C.-V. López-Torres, S. Salazar Colores, K. Kells, J.-C. Pedraza-Ortega, and J.-M. Ramos-Arreguin, "Improving 3d reconstruction accuracy in wavelet transform profilometry by reducing shadow effects," *IET Image Processing*, vol. 14, no. 2, pp. 310–317, 2020
16. Hosseinzadeh, S., Shakeri, M., Zhang, H.: "Fast shadow detection from a single image using a patched convolutional neural network,"in. IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS) **2018**, 3124–3129 (2018)

17. B. Ding, C. Long, L. Zhang, and C. Xiao, "Argan: Attentive recurrent generative adversarial network for shadow detection and removal," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 10 213–10 222
18. Wang, C., Pang, Q.: The elimination of errors caused by shadow in fringe projection profilometry based on deep learning. Opt. Lasers Eng. **159**, 107203 (2022)
19. Mantravadi, M., Malacara, D.: Newton, fizeau, and haidinger interferometers. Optical shop testing **59**, 1–45 (1992)
20. Zuo, C., Huang, L., Zhang, M., Chen, Q., Asundi, A.: Temporal phase unwrapping algorithms for fringe projection profilometry: A comparative review. Opt. Lasers Eng. **85**, 84–103 (2016)
21. Takeda, M., Mutoh, K.: Fourier transform profilometry for the automatic measurement of 3-d object shapes. Appl. Opt. **22**(24), 3977–3982 (1983)
22. Feng, S., Zuo, C., Zhang, L., Tao, T., Hu, Y., Yin, W., Qian, J., Chen, Q.: Calibration of fringe projection profilometry: A comparative review. Opt. Lasers Eng. **143**, 106622 (2021)
23. Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 10 012–10 022
24. Ravi, V., Gorthi, R.K.: Litef2dnet: a lightweight learning framework for 3d reconstruction using fringe projection profilometry. Appl. Opt. **62**(12), 3215–3224 (2023)
25. Sumanth, K., Ravi, V., Gorthi, R.K.: A multi-task learning for 2d phase unwrapping in fringe projection. IEEE Signal Process. Lett. **29**, 797–801 (2022)
26. Ravi, V., Gorthi, R.K.: Cf3dnet: A learning-based approach for single-shot 3d reconstruction from circular fringes. Opt. Lasers Eng. **167**, 107597 (2023)
27. Y. Zheng, S. Wang, Q. Li, and B. Li, "Fringe projection profilometry by conducting deep learning from its digital twin," *Optics express*, vol. 28, no. 24, pp. 36 568–36 583, 2020
28. Kemao, Q.: Windowed fourier transform for fringe pattern analysis. Appl. Opt. **43**(13), 2695–2702 (2004)
29. Zhao, M., Huang, L., Zhang, Q., Su, X., Asundi, A., Kemao, Q.: Quality-guided phase unwrapping technique: comparison of quality maps and guiding strategies. Appl. Opt. **50**(33), 6214–6224 (2011)

EMPATH: MediaPipe-Aided Ensemble Learning with Attention-Based Transformers for Accurate Recognition of Bangla Word-Level Sign Language

Kazi Reyazul Hasan and Muhammad Abdullah Adnan(✉)

Department of Computer Science and Engineering, Bangladesh University of Engineering and Technology, Dhaka, Bangladesh
abdullah.adnan@gmail.com

Abstract. In this paper, we introduce EMPATH, an advanced computational framework developed to substantially enhance the recognition of Bangla Sign Language (BdSL). By integrating **E**nsemble Learning, **M**edia**P**ipe Holistic for gesture tracking, and an **A**ttention-based **T**ransformer model, EMPATH addresses the challenges of sign language interpretation, setting new accuracy benchmarks and significantly surpassing previous records: achieving 79.81% on SignBD-Word-90 (previously best at 66.05%), 70.58% on SignBD-Word (previously best at 57%), and 99.25% on BdSL40 (previously best at 89%).

A pioneering feature of EMPATH is its innovative interpolation model, built to overcome the limitations posed by missing **H**and keypoints of MediaPipe. Validated in both EMPATH and a basic MLP framework, this model showcases remarkable versatility across architectures. EMPATH strategically selects its preprocessing and postprocessing techniques, optimizing each for maximum impact on accuracy and performance.

Extensively trained across various word-level datasets beyond Bangla, including INCLUDE-50, INCLUDE, WLASL-100, and the Malaysian Sign Language Medical Dataset, EMPATH demonstrates its broad applicability and global potential. This diverse training scheme establishes superior accuracy benchmarks: achieving an impressive 100% on INCLUDE-50, 94.67% on INCLUDE, and 93.46% on the MSL Medical Dataset.

Through EMPATH, we aspire to bridge communication barriers for the deaf and hard-of-hearing communities, showcasing the profound impact of integrating advanced technological solutions to tackle the complexities of sign language recognition.

Source code is available at https://github.com/kreyazulh/EMPATH.

Keywords: Gesture Recognition · Sign Language Interpretation · Keypoints Detection

A. Antonacopoulos et al. (Eds.): ICPR 2024, LNCS 15321, pp. 355–371, 2025.
https://doi.org/10.1007/978-3-031-78305-0_23

1 Introduction

Sign language recognition is a bridge for the deaf and hard-of-hearing communities, enabling essential communication and promoting inclusivity. Bangla Sign Language (BdSL), significant yet underrepresented in computational recognition, faces challenges due to the lack of proper datasets and continuation of research on improvement. Previous efforts have focused more on dataset creation than on advancing recognition technologies, particularly for BdSL, leaving a gap in subsequent research and application.

EMPATH emerges as a comprehensive framework, making a significant leap in BdSL recognition. It integrates Ensemble Learning [8], MediaPipe Holistic [14] for gesture tracking, and Attention-based Transformer models [23] as shown in Fig. 1, tackling the complex challenges of sign language interpretation with a synergistic approach. Each component is chosen for its proven effectiveness, with Ensemble Learning boosting reliability, MediaPipe providing advanced tracking, and Transformers enhancing context processing.

The innovation of EMPATH is showcased in its pioneering application to Bangla word-level sign language datasets, leveraging advanced modeling and state-of-the-art augmentation techniques like frame dropping, time-scaling, and quantization. Central to EMPATH's advancements is a novel interpolation model specifically designed to address the challenge of missing hand keypoints—a notable limitation with MediaPipe technologies. This model has been rigorously validated across both the comprehensive EMPATH framework and a straightforward Multilayer Perceptron (MLP) model [21], affirming its widespread applicability.

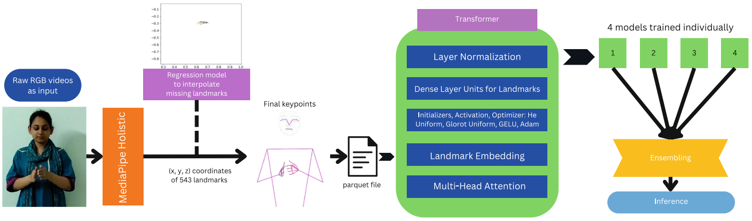

Raw RGB videos as input
MediaPipe Holistic
Regression model to interpolate missing landmarks
(x, y, z) coordinates of 543 landmarks
Final keypoints
parquet file
Transformer
Layer Normalization
Dense Layer Units for Landmarks
Initializers, Activation, Optimizer: He Uniform, Glorot Uniform, GELU, Adam
Landmark Embedding
Multi-Head Attention
4 models trained individually
1
2
3
4
Ensembling
Inference

Fig. 1. High-level Overview of EMPATH: The process begins with raw RGB videos fed into MediaPipe Holistic, extracting landmarks. Should any hand keypoints be missing, our interpolation model predicts and fills in the gaps. The data, now complete, is converted to a parquet file format for efficient processing. Subsequently, the Transformer model, initialized with parameters like Layer Normalization and Landmark Embedding, takes over. The architecture employs Multi-Head Attention alongside dense layers activated by GELU functions. Upon Transformer training completion, the results from four individually trained models are aggregated using an ensembling method. This collaborative decision-making leads to the final inference, resulting in the accurate interpretation of sign language.

A distinctive aspect of our approach is the custom-built Transformer model, coded entirely from scratch to ensure full control over parameters. The model's lightweight architecture for the use of TFLite makes it suitable for deployment on lower-end devices, such as mobile phones, broadening the potential for real-world application and accessibility.

Furthermore, EMPATH enhances its predictive accuracy through ensemble modeling, a technique that involves training four separate Transformer models and employing majority voting for decision-making.

EMPATH has notably enhanced recognition on BdSL word level datasets, achieving 79.81% on SignBD-Word-90, 69.42% on SignBD-Word [17], and 99.25% on BdSL40 [19], setting new benchmarks. Its adaptability extends to other languages, evidenced by its success on datasets like INCLUDE-50, INCLUDE [20], and the Malaysian Sign Language Medical Dataset [1], achieving 100%, 94.67%, and 93.46%, respectively. These achievements highlight EMPATH's role in reducing communication barriers and its potential for universal sign language recognition.

EMPATH along with our other proposed techniques have been rigorously trained, tested and validated across multiple settings, highlighting their strong performance and pinpointing opportunities for refinement. Our contribution is summarized as follows:

- We introduce EMPATH, a comprehensive framework integrating Ensemble Learning, MediaPipe Holistic, and Attention-based Transformer models for improved sign language recognition.
- We develop a novel interpolation model to address the challenge of missing hand keypoints, enhancing the accuracy of sign interpretation.
- We custom-build a lightweight Transformer model from scratch, optimized for deployment on low-end devices to increase accessibility.
- We set new benchmarks not only in Bangla Sign Language (BdSL) recognition but also across a variety of datasets, showcasing the effectiveness and adaptability of EMPATH to different languages.

Our paper is structured sequentially, covering sections dedicated to related works, methodology, experiments, and a concluding segment.

2 Related Works

Research in sign language recognition has extensively utilized datasets to advance the understanding of various sign languages. Predominantly, American Sign Language (ASL) datasets like Purdue RVL-SLLL ASL [24], WLASL [13], and MS-ASL [26] have been pivotal, providing vast numbers of RGB videos from multiple signers. These resources have been instrumental in developing sophisticated recognition techniques. Other significant contributions include LSA64 [16] for Argentinean Sign Language, capturing unique regional signs, and DEVISIGN [7] for Chinese Sign Language, both reflecting the linguistic diversity within the sign language research community. For comprehensive details on the datasets

Table 1. An overview of various word-level sign language video datasets. The datasets highlighted are those that have been worked on in this paper.

Datasets	Language	Classes	Videos	Signers
Purdue RVL-SLLL ASL [24]	American	104	2576	14
WLASL [13]	American	2000	21083	119
MS-ASL [26]	American	1000	9764	222
LSA64 [16]	Argentinian	64	3200	10
BdSL-40 (INCLUDE Subset) [19]	Bangla	40	611	7
SignBD-Word [17]	Bangla	200	6000	16
BdSLW-60 [3]	Bangla	60	9307	18
DEVISIGN [7]	Chinese	2000	24000	8
INCLUDE [20]	Indian	263	4287	7
MSL Medical Dataset [1]	Malaysian	50	1040	7

discussed, as well as others, refer to Table 1. This includes well-known datasets and those we intend to explore further. Additionally, numerous other datasets are available for various sign languages, contributing to the broader field of sign language research.

In contrast, research on Bangla Sign Language (BdSL) remains limited, particularly at the word level. While there is a reasonable foundation of alphabet-level BdSL datasets [11,15], there has been a notable dearth of word-level data, which is essential for comprehensive communication. The BdSL-40, an adaptation from the INCLUDE dataset, the SignBD-Word, the BdSLW-60 [3] datsets are some of the recent endeavors towards creating word-level BdSL datasets. Although these are commendable efforts, they have not been extensively explored or validated in subsequent research, leaving a vast area for potential exploration and development.

Existing work on BdSL has not utilized the full potential of advanced computational models. For instance, while MediaPipe has been adopted for gesture recognition in various domains, its integration into BdSL recognition [10] has been overlooked or alternative approaches [2] have been employed, and no work addressing the challenge of missing hand keypoints has been presented. These keypoints are critical for accurate gesture interpretation, and their absence can significantly diminish a model's performance.

Research into sign language recognition has explored a range of computational models, reflecting the diversity and complexity of interpreting sign languages. Classical approaches have often relied on convolutional neural networks (CNNs) [22], given their proficiency in processing visual data. Recurrent neural networks, particularly Long Short-Term Memory (LSTM) [3,12,17,20] networks, have also been extensively utilized due to their capability to capture temporal sequences which are inherent in sign language data. Graph Neural Networks (GNNs) [25], Inflated 3D ConvNets (I3D) [6] and Visual-Language Pretraining (VLP) [27] have also emerged as alternatives, capitalizing on their ability to model intricate spatial structures and temporal information respectively.

However, the adoption of Transformer models in sign language recognition has been limited. Despite their success in various machine learning domains due to their sophisticated handling of sequential data and long-range dependencies—their full potential in sign language recognition remains untapped, specially in

BdSL works. There have been instances where Transformers have been applied and shown promising results [5].

Ensemble modeling is another area that has not been fully explored within the context of sign language recognition. While the technique is well-known for its ability to improve performance by combining the strengths of multiple models through strategies like majority voting or stacking, its application has been scant in comparison to the individual model architectures.

In light of these gaps, our work aims to advance the field by introducing a complete approach that synergistically combines MediaPipe, Transformer models, and ensemble learning. Our method addresses the critical issue of missing keypoints in MediaPipe and proposes a solution to enhance accuracy and robustness. By benchmarking our approach on various datasets and providing comparisons with existing works, we illuminate the path forward for future research in this domain.

3 Methodology

3.1 Data Preparation and Landmark Extraction

Data Preparation: The initial step involves utilizing RGB videos from the dataset to our research and systematically organizing them. Each video is tagged with specific labels that represent the conveyed sign language gestures.

MediaPipe Holistic for Landmark Extraction: Following the organization of videos and labels, we employ MediaPipe Holistic for extracting pose, hand, and face landmarks. MediaPipe Holistic is part of the MediaPipe Framework [14], a comprehensive toolkit designed for constructing efficient machine learning pipelines on devices, enabling real-time sign language recognition. MediaPipe Holistic is configured for landmark detection and tracking with minimum detection and tracking confidence set at 0.5 and model complexity at 1. Although setting the model complexity to 2 could potentially enhance performance, it significantly extends inference time. The selection of tracking and detection parameters underwent thorough testing to ensure optimal performance.

Extracting Landmarks: With the MediaPipe Holistic configured, each video undergoes processing to extract landmarks associated with poses, hands, and faces. Post-extraction, the landmarks are stored in Parquet files corresponding to each video, facilitating efficient data management and accessibility for future analysis and model training. The choice of Parquet files is due to their compact data structure and rapid read/write capabilities, making them highly suitable for managing the voluminous datasets typical in sign language recognition research.

3.2 Transformer Training

Hyperparameters Setting and Preprocessing: Our training setup is defined by several critical parameters: we opt for training with or without preprocessing data (`PREPROCESS_DATA`), and setting specific hyperparameters like input size (`INPUT_SIZE = 12 or 24`), batch size (`BATCH_SIZE = 16`), and learning rate (`LR_MAX = 1e-3`). The model is trained over 300 epochs (`N_EPOCHS = 300`) with a warm-up period (`N_WARMUP_EPOCHS = 0`). We employ a weight decay ratio (`WD_RATIO = 0.05`) and use a masking value for data padding. We also introduce a preprocessing layer designed for Tensorflow to handle data within the TFLite environment, where data preprocessing must occur within the model due to the inability to utilize Python for such tasks. This layer, `PreprocessLayer`, applies essential preprocessing steps, including padding, downsampling, and filtering frames with insufficient hand data.

Model Configuration and Transformer Architecture: Our transformer model is designed to address the intricate challenges of sign language recognition, focusing on the effective processing of landmark data. The model configuration combines embedding layers with transformer blocks to capture the dynamic spatial and temporal information present in sign language videos.

The core elements of our model configuration are as follows:

1. **Layer Normalization:** We apply layer normalization with an epsilon value set to 1×10^{-6} (`LAYER_NORM_EPS`), enhancing the stability of the learning process.
2. **Dense Layer Units for Landmarks:** Dedicated dense layers for lips, hands, and pose (`LIPS_UNITS, HANDS_UNITS, POSE_UNITS` $= 384$) converge into a final embedding size (`UNITS` $= 512$).
3. **Transformer Architecture:** The architecture comprises three transformer blocks (`NUM_BLOCKS` $= 3$) with an MLP ratio of 3 (`MLP_RATIO`), utilizing scaled dot-product attention [23]:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

   where Q, K, and V are the queries, keys, and values, respectively, and d_k denotes the dimensionality of the keys.
4. **Dropout:** To combat overfitting, dropout is strategically applied during embedding (`EMBEDDING_DROPOUT` $=$ 0.00), in MLP layers (`MLP_DROPOUT_RATIO` $= 0.30$), and at the classifier stage (`CLASSIFIER_DROPOUT_RATIO` $= 0.10$).
5. **Initializers and Activations:** The He uniform and Glorot uniform initializers, along with GELU activation functions, are employed to optimize the model's training efficiency.
6. **Landmark Embedding:** Custom embeddings for lips, hands, and pose landmarks are designed to adeptly encode spatial features, incorporating an empty

embedding vector for frames with missing landmarks, initialized to zeros (full body pose is always not covered in videos). Hand landmarks were given a higher weight while embedding due to its greater significance on sign expressions.

7. **Multi-Head Attention:** The model leverages multi-head attention to simultaneously focus on various segments of the input sequence, enhancing its capability to attend complex dependencies. The overall model structure is depicted in Fig. 2

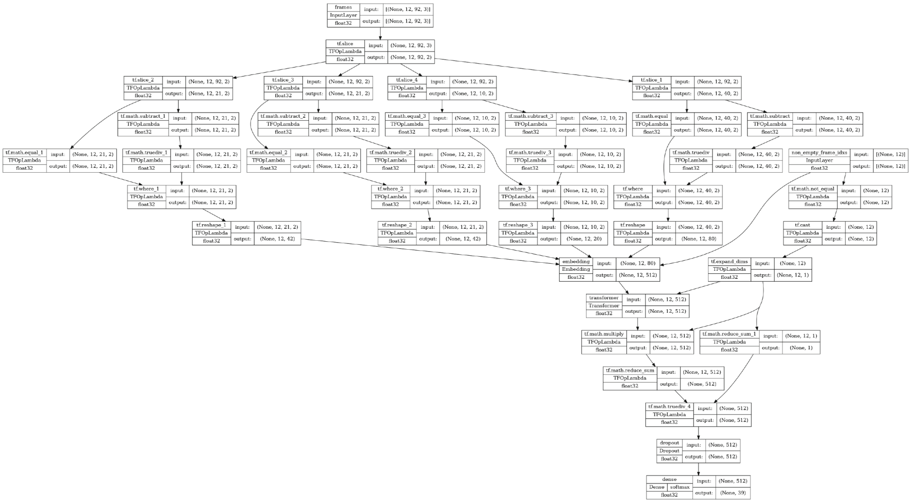


Fig. 2. Transformer Architecture with Multi-branch Parallel Processing

This architecture is crafted to adeptly process and interpret the challenging patterns of sign language, providing a solid foundation for achieving accurate and reliable sign language recognition.

3.3 Augmentation Techniques and Postprocessing

Augmentation Techniques: To enhance the generalizability of our model, we employed two key augmentation techniques: frame drop and time-scale augmentation. These methods introduce variability into the training process.

Frame Drop Augmentation probabilistically omits certain frames from the video sequences. This approach simulates scenarios where frames might be missing due to recording errors or transmission issues, training the model to maintain performance even when input data is incomplete.

Time-Scale Augmentation modifies the temporal dimension of the video data by either compressing or expanding the time scale of certain sequences.

This simulates variations in sign execution speed among different individuals, accommodating for the natural diversity in sign language delivery.

While these augmentation techniques introduce beneficial variability, larger values or probabilities can reduce accuracy. Therefore, it is crucial to test and determine optimal values to ensure the model's performance is not adversely affected.

Model Quantization: We apply quantization to reduce the model's precision from 32-bit floating-point to either 16-bit floating-point (`fp16`) or 8-bit integers (`int8`), depending on the deployment needs. This results in a significant reduction in model size, typically down to approximately 60-70 MB. This reduction simplifies deployment on mobile and embedded devices, where storage and computational resources are typically constrained.

Without quantization, the larger 32-bit floating-point values make models impractical for smaller devices, and their higher precision often doesn't noticeably improve accuracy.

3.4 Ensemble Technique

We employed an ensemble technique using four distinct transformer models (`N_MODELS = 4`) rather than a single model for increasing overall accuracy. Each model is trained in the same environment. Our augmentation techniques are probabilistic for each batch frames, introducing variation and different training schemes for ensemble modeling. During inference, we aggregated the predictions of these four models using majority voting, meaning the final output is based on the prediction that the majority of the models agree upon. This strategy reduces the likelihood of individual incorrect predictions due to very close prediction scores among some labels.

We chose four models to balance accuracy and computational efficiency (see table 2).

Table 2. Validation of the number of transformer models chosen for the ensemble method

Dataset	Trained:1	Trained:2	Trained:3	Trained:4	Trained:5
SignBD-Word-90	76.67%	76.67%	78.51%	79.81%	79.81%
INCLUDE-50	97.93%	98.44%	100%	100%	100%

Training multiple models takes time, and the final model size also increases. Hence, the choice of the number of models was made considering these tradeoffs. We observed that for the datasets we worked with, the ensemble technique did not increase accuracy after training four models. While three models also provided good performance, the accuracy for SignBD-Word-90 increased slightly with four models. Therefore, we decided to use four models for our ensemble technique. This choice optimizes the trade-off between accuracy, training time, and model size.

3.5 Interpolation Model for Missing Keypoints

To address the challenge of missing hand keypoints in sign language video frames, we apply a linear interpolation model. This is a preprocessing technique, utilized when extracting keypoints from MediaPipe. While acknowledging that human motion's complexity may not be fully captured by a linear model, the relative continuity of motion allows for effective linear approximations. This method involves identifying frames before and after the missing sequence where keypoints are fully present. Using these as reference, we interpolate the values for missing frames based on linear progression.

Given hx as the x-coordinate of a hand landmark, $last_{hx}$ and $next_{hx}$ represent the x-coordinates of the last and next complete frames, respectively. For a missing frame positioned at M_x within a gap of $N_x - L_x$ frames, the predicted x-coordinate is calculated as:

$$predicted_{hx} = last_{hx} + (next_{hx} - last_{hx})\frac{M_x}{N_x - L_x}$$

This process is replicated for y and z coordinates across all 22 hand landmarks to generate a complete set of predicted values.

An illustrative example of generating missing values is presented in Fig. 3.

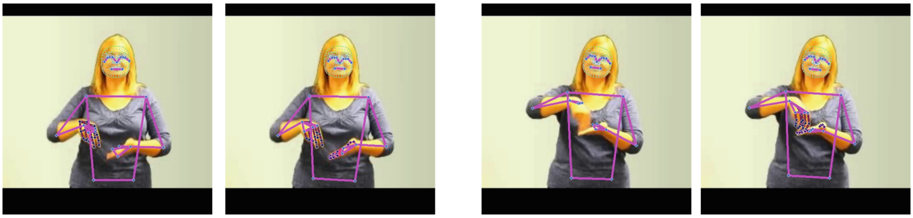


Fig. 3. Missing keypoints regrenerated and arm aligned using the interpolation model (the right image in each set). Frames are taken from a video of WLASL.

We also employ a transformation that aligns the pose wrist coordinates with the hand wrist coordinates, considering the wrist as a reliable root detection point. The translation is computed as the difference in coordinates between the hand and pose wrist, utilizing homogeneous coordinates for transformation:

$$t_x = hand_x - pose_x; \quad t_y = hand_y - pose_y; \quad t_z = hand_z - pose_z;$$

The transformation matrix and the original wrist coordinates are then multiplied to obtain the new aligned pose wrist coordinates:

$$\begin{bmatrix} 1 & 0 & 0 & t_x \\ 0 & 1 & 0 & t_y \\ 0 & 0 & 1 & t_z \\ 0 & 0 & 0 & 1 \end{bmatrix} \times \begin{bmatrix} x \\ y \\ z \\ 1 \end{bmatrix} = \begin{bmatrix} x' \\ y' \\ z' \\ 1 \end{bmatrix}$$

Algorithm 1 Hand Landmarks Prediction

```
Require: Video input file
Ensure: Processed video with improved landmarks
Initialize video capture, holistic model, variables and batch size
while video frames available do
    Read and resize frame size if required
    Process the frame using the holistic model and store frame info
    Identify and update missing hand landmarks
    if batch size reached or end of video then
        Calculate missing landmarks using adjacent frames
        Draw fixed landmarks on frames and output
        Clear temporary lists to free memory
    end if
end while
Release resources and close model
```

We employ the `batch size` variable to manage memory usage and prevent exhaustion. We noticed a marginal increase of only 1.83% in average in batch process using the interpolation model (see supplementary material section 4).

4 Experiments

In the experimental section, we evaluate EMPATH across multiple datasets, including BdSL datasets, INCLUDE, WLASL, and the MSL Medical dataset. The focus is to assess the models' generalization capabilities and their effectiveness in sign language recognition. Additionally, we test a fundamental MLP model with our interpolation technique for handling missing keypoints, contrasting its generalization capacity along with EMPATH's complex framework. Our experiments were conducted on a computational environment powered by a P100 GPU.

Evaluation Metrics We employ accuracy (top-1 and top-5) as our primary metric to compare the outcomes against existing benchmarks. Confusion matrix is deployed for visualization where it seems fit. We also incorporate macro-averaged F1 score into our evaluation in specific cases.

4.1 BdSL Datasets

SignBD-Word: The SignBD-Word dataset is a substantial video-based resource for word-level Bangla Sign Language (BdSL) research, publicly available and comprising over five hours of footage. It features 16 signers each performing 200 unique signs, totaling 6000 instances of sign language data. SignBD-Word also offers a subset of 90 words, known as SignBD-Word-90. Our EMPATH model, tested on both sets following the source train-test split, showcased remarkable performance. It achieved a top-1 accuracy of 79.81% on SignBD-Word-90 and an impressive 70.58% on SignBD-Word, outperforming established benchmarks by substantial margins (see Table 3).

Table 3. Top-1 and Top-5 accuracy(%) of different models on SignBD-Word

Model	**SignBD-Word-90**		**SignBD-Word**	
	Top-1	**Top-5**	**Top-1**	**Top-5**
CNN-LSTM [17]	10.8	31.92	9.62	28.45
CNN-Attention-LSTM [17]	21.5	39.25	15.6	30.56
I3D [6]	64.2	88.25	52.5	80
SlowFast [9]	66.05	88.82	57	84.17
EMPATH (ours)	**79.81**	**93.15**	**70.58**	**88.08**

The original source dataset contained two versions: one as RGB videos and another with skeletal keypoints. Our approach was to process the original RGB videos according to the needs of the EMPATH model.

BdSL-40: The BdSL-40 dataset serves as a valuable resource in Bangla Sign Language recognition, featuring 611 videos across 40 words. Stemming directly from INCLUDE, it utilizes the linguistic affinity with West Bengal and Indian sign languages. The adaptation process for creating the BdSL-40 dataset from INCLUDE involved a review of the Bangladesh Sign Language Dictionary to identify corresponding signs. We applied our EMPATH model to BdSL-40 and accomplished a remarkable accuracy of 99.25% (see Table 4, Fig. 4), setting an almost perfect benchmark. Notably, the original dataset comprised 39 classes and 603 videos when downloaded from source.

We have incorporated some challenging label frames from BdSL-40 in the example shown in Fig. 5. These labels are challenging because they have many similarities in sign actions, making them difficult to predict accurately for less powerful models or optimization techniques.

4.2 Other Language Datasets

INCLUDE: Transitioning to the INCLUDE dataset, we applied our EMPATH model, which excelled on BdSL-40, to this broader lexicon of Indian Sign Language (ISL). INCLUDE features a total of 4,287 videos, spanning 263 word signs

Table 4. Comparison of model accuracies on BdSL-40

Model	Accuracy (%)
3D-CNN [22]	82.43
GNN [25]	89.00
EMPATH (Ours)	**99.25**

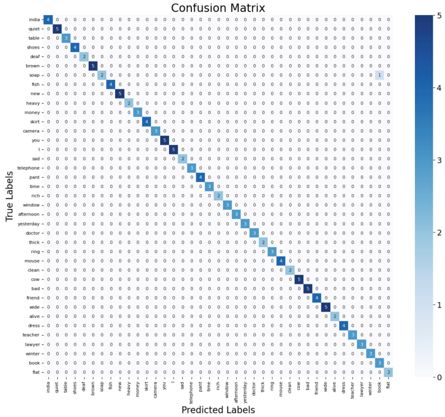


Fig. 4. BdSL-40

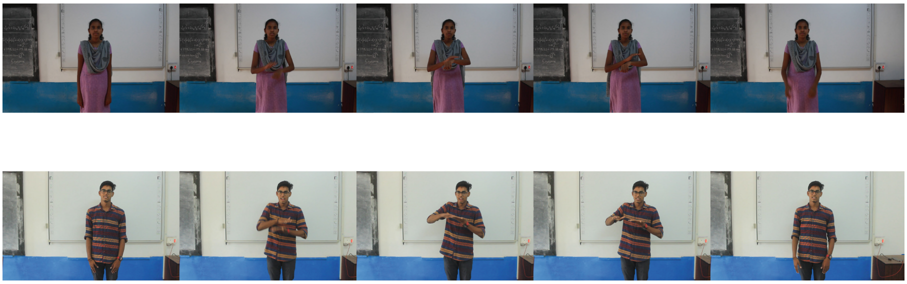


Fig. 5. First row: Video frames with ground truth "soap". Second row: Video frames with ground truth "flat". These labels were interchangeably predicted when tested with a base MLP model and a single transformer model without preprocessing, ensembling, and augmentation techniques. However, with the full integration of the EMPATH system, these labels have been successfully identified and predicted accurately.

categorized into 15 different classes. INCLUDE-50, a subset featuring 50 word signs across categories, is used to fine-tune model parameters. Adhering to the original train-test split from the source—correcting minor dataset discrepancies (see supplementary material section 5)—we achieved remarkable results. We set new benchmarks: a perfect 100% accuracy on INCLUDE-50 (see Table 5 and Fig. 6) and 94.67% on the complete INCLUDE dataset, reinforcing the power and versatility of EMPATH.

While employing the INCLUDE dataset for our EMPATH model, we noted the absence of the 'second' (number) class in the original data. Consequently, we proceeded with the evaluation using 262 classes (see Table 6 and Fig. 7).

WLASL-100: The WLASL dataset, an extensive compilation of American Sign Language (ASL), consists of 2000 classes that contribute significantly to sign language research. Our focus was on the WLASL-100 subset, which ideally contains

Table 5. Accuracy of different methods on INCLUDE-50

Model	Accuracy (%)
XGBoost [20]	89.1
MobileNetV2+BiLSTM [20]	94.5
MediaPipe+LSTM [?]	94.8
EMPATH (Ours)	**100**

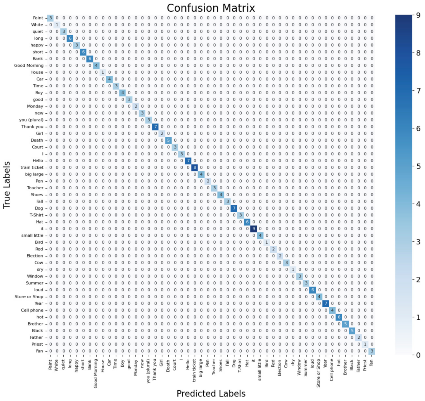


Fig. 6. INCLUDE-50

Table 6. Accuracy of different methods on INCLUDE

Model	Accuracy (%)
XGBoost [20]	61.1
MobileNetV2+BiLSTM [20]	85.6
MediaPipe+LSTM [12]	87.4
SL-GCN [18]	93.5
EMPATH (Ours)	**94.67**

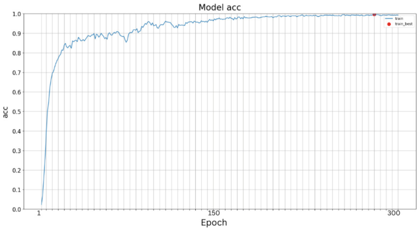


Fig. 7. Training Accuracy per Epoch

over 2000 videos. However, due to broken links and missing content, we initially retrieved only 848 files, averaging fewer than seven videos per class to train. Despite the challenges, we secured the remaining videos to complete the dataset. Nonetheless, we chose to test the generalization ability of our EMPATH model on the limited dataset to understand its performance under data-constrained conditions. With only a fraction of the full dataset, EMPATH achieved a top-1 accuracy of 56.25%, a top-5 accuracy of 78.91%, and a macro-averaged F1 score of 52.63%. In comparison, models like I3D [6] and SPOTER [4] reached top-1 accuracies of 65.89% and 63.18%, respectively, using the complete dataset without pretraining. EMPATH's performance, with less than half the training set, showcases its robust class generalization capabilities, underscoring the experiment's aim to demonstrate its efficiency in data-limited scenarios.

MSL Medical Dataset: The MSL Medical Dataset represents the final dataset in our series of evaluations, and its selection was strategic due to its structure and unestablished benchmarks. This dataset uniquely encompasses both word and sentence classes, with sentences incorporating words that are also categorized independently. Such a structure allowed us to go deeper into EMPATH's generalization abilities. From this dataset, we extracted a total of 49 classes,

Table 7. EMPATH model performance on MSL Medical Dataset

Metric	Score (%)
Top-1 Accuracy	93.46
Top-5 Accuracy	99.53
Macro F1 Score	92.35

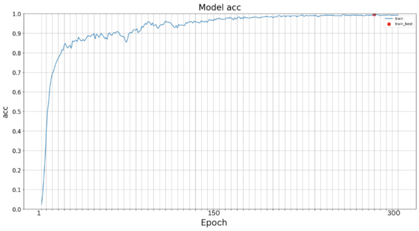


Fig. 8. Challenge in MSL Medical Dataset: EMPATH struggled when context is present in both word and sentence class, incorrectly interpreting "body" in sentences like "my body feels itchy", leading to three incorrect predictions

inclusive of sentence labels. Our findings were compelling—EMPATH reached a top-1 accuracy of 93.46%, a top-5 accuracy of 99.53%, and a macro-averaged F1 score of 92.35% (see Table 7), thereby underscoring the model's capacity for class generalization, even within datasets rich in linguistic complexity.
Fig. 8 illustrates a rare instance of EMPATH's failure to generalize. However, this is not the common case as the model successfully interprets instances like the word "tired" and the sentence "I'm vomiting and tired because I am pregnant."

4.3 Interpolation Model for Handling MediaPipe Missing Keypoints

Concluding our experimental section, we reflect on the effectiveness of the interpolation model employed as a preprocessing technique to handle missing keypoints from MediaPipe's output. This model was utilized across all our prior experiments. By incorporating the interpolation model, we sought to portray its impact on two training approaches: the advanced EMPATH model and the fundamental MLP model. Our tests, conducted on the BdSL-40 and MSL Medical Dataset, compared the performance outcomes with and without interpolation preprocessing. The results (see Table 8) from this inquiry were intended to validate the interpolation model's wide-ranging utility and its adaptability to an array of machine learning architectures.

Table 8. Performance comparison with and without interpolation preprocessing

Model	Dataset	Accuracy (no interpolation)	Accuracy (with interpolation)
EMPATH	BdSL-40	97.02%	**99.25%**
MLP	BdSL-40	77.61%	**79.10%**
EMPATH	MSL Medical	91.16%	**93.46%**
MLP	MSL Medical	66.97%	**69.30%**

In summary, we evaluated our EMPATH model across multiple sign language datasets, set new performance benchmarks in sign language recognition, tested our proposed interpolation model for handling missing keypoints for acceptance across multiple architectures. Detailed ablation studies on augmentation techniques can be found in supplementary material section 6.

5 Conclusion

The EMPATH framework has notably enhanced sign language recognition, incorporating MediaPipe Holistic and Transformer models to achieve breakthrough performance on diverse datasets. Despite its success, EMPATH encounters difficulties with overlapping contexts and in handling missing keypoints in low-quality videos due to simple interpolation technique implementation. Future efforts will focus on refining context recognition, advancing keypoint reconstruction techniques, and incorporating non-manual signals for richer interpretations.

Acknowledgements. The research benefited from partial support provided by the ICT Innovation Fund, Government of Bangladesh, and the RISE Student Research Grant from Bangladesh University of Engineering and Technology. The development of the EMPATH model was inspired by collaborative insights gained during the Google - Isolated Sign Language Recognition competition on Kaggle.

References

1. Abuan, A.V., Rahman, M.Z., Abuan, A.D., Lee, S.H.: Malaysian Sign Language Medical Dataset (Jul 2023), https://github.com/Arekku21/MSL-Medical
2. Akash, S.K., Chakraborty, D., Kaushik, M.M., Babu, B.S., Zishan, M.S.R.: Action recognition based real-time bangla sign language detection and sentence formation. In: 2023 3rd International Conference on Robotics, Electrical and Signal Processing Techniques (ICREST). pp. 311–315. IEEE (2023)
3. Ara Rubaiyeat, H., Mahmud, H., Habib, A., Kamrul Hasan, M.: Bdslw60: A word-level bangla sign language dataset. arXiv e-prints pp. arXiv–2402 (2024)
4. Boháček, M., Hrúz, M.: Sign pose-based transformer for word-level sign language recognition. In: Proceedings of the IEEE/CVF winter conference on applications of computer vision. pp. 182–191 (2022)
5. Camgoz, N.C., Koller, O., Hadfield, S., Bowden, R.: Sign language transformers: Joint end-to-end sign language recognition and translation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10023–10033 (2020)
6. Carreira, J., Zisserman, A.: Quo vadis, action recognition? a new model and the kinetics dataset. In: proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 6299–6308 (2017)
7. Chai, X., Wang, H., Chen, X.: The devisign large vocabulary of chinese sign language database and baseline evaluations. In: Technical report VIPL-TR-14-SLR-001. Key Lab of Intelligent Information Processing of Chinese Academy of Sciences (CAS). Institute of Computing Technology (2014)
8. Dietterich, T.G.: Ensemble methods in machine learning. In: International workshop on multiple classifier systems. pp. 1–15. Springer (2000)
9. Feichtenhofer, C., Fan, H., Malik, J., He, K.: Slowfast networks for video recognition. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 6202–6211 (2019)

10. Foysol, M.W., Sajal, S.E.A., Alam, M.J.: Vision-based real time bangla sign language recognition system using mediapipe holistic and lstm. In: 2023 International Conference on Computational Intelligence and Knowledge Economy (ICCIKE). pp. 19–24. IEEE (2023)
11. Hoque, O.B., Jubair, M.I., Akash, A.F., Islam, S.: Bdsl36: A dataset for bangladeshi sign letters recognition. In: Proceedings of the Asian Conference on Computer Vision (2020)
12. Khartheesvar, G., Kumar, M., Yadav, A.K., Yadav, D.: Automatic indian sign language recognition using mediapipe holistic and lstm network. Multimedia Tools and Applications **83**(20), 58329–58348 (2024)
13. Li, D., Rodriguez, C., Yu, X., Li, H.: Word-level deep sign language recognition from video: A new large-scale dataset and methods comparison. In: Proceedings of the IEEE/CVF winter conference on applications of computer vision. pp. 1459–1469 (2020)
14. Lugaresi, C., Tang, J., Nash, H., McClanahan, C., Uboweja, E., Hays, M., Zhang, F., Chang, C.L., Yong, M.G., Lee, J., et al.: Mediapipe: A framework for building perception pipelines. arXiv preprint arXiv:1906.08172 (2019)
15. Podder, K.K., Chowdhury, M., Mahbub, Z.B., Kadir, M.: Bangla sign language alphabet recognition using transfer learning based convolutional neural network. Bangladesh J. Sci. Res pp. 31–33 (2020)
16. Ronchetti, F., Quiroga, F.M., Estrebou, C., Lanzarini, L., Rosete, A.: Lsa64: an argentinian sign language dataset. arXiv preprint arXiv:2310.17429 (2023)
17. Sams, A., Akash, A.H., Rahman, S.M.: Signbd-word: Video-based bangla word-level sign language and pose translation. In: 2023 14th International Conference on Computing Communication and Networking Technologies (ICCCNT). pp. 1–7. IEEE (2023)
18. Selvaraj, P., Nc, G., Kumar, P., Khapra, M.: Openhands: Making sign language recognition accessible with pose-based pretrained models across languages. arXiv preprint arXiv:2110.05877 (2021)
19. Shahgir, H.S., Sayeed, K.S., Tahmid, M.T., Zaman, T.A., Alam, M.Z.U.: Connecting the dots: Leveraging spatio-temporal graph neural networks for accurate bangla sign language recognition. arXiv preprint arXiv:2401.12210 (2024)
20. Sridhar, A., Ganesan, R.G., Kumar, P., Khapra, M.: Include: A large scale dataset for indian sign language recognition. In: Proceedings of the 28th ACM international conference on multimedia. pp. 1366–1375 (2020)
21. Taud, H., Mas, J.F.: Multilayer perceptron (mlp). Geomatic approaches for modeling land change scenarios pp. 451–455 (2018)
22. Tran, D., Bourdev, L., Fergus, R., Torresani, L., Paluri, M.: Learning spatiotemporal features with 3d convolutional networks. In: Proceedings of the IEEE international conference on computer vision. pp. 4489–4497 (2015)
23. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in neural information processing systems **30** (2017)
24. Wilbur, R., Kak, A.C.: Purdue rvl-slll american sign language database (2006)
25. Yu, B., Yin, H., Zhu, Z.: Spatio-temporal graph convolutional networks: A deep learning framework for traffic forecasting. arXiv preprint arXiv:1709.04875 (2017)
26. Zahedi, M., Dreuw, P., Rybach, D., Deselaers, T., Bungeroth, J., Ney, H.: Continuous sign language recognition–approaches from speech recognition and available data resources. In: sign-lang@ LREC 2006. pp. 21–24. European Language Resources Association (ELRA) (2006)

27. Zhou, B., Chen, Z., Clapés, A., Wan, J., Liang, Y., Escalera, S., Lei, Z., Zhang, D.: Gloss-free sign language translation: Improving from visual-language pretraining. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 20871–20881 (2023)

Cross-Attention Based Influence Model for Manual and Nonmanual Sign Language Analysis

Lipisha Chaudhary(✉), Fei Xu, and Ifeoma Nwogu

Department of Computer Science and Engineering, University at Buffalo,
Buffalo, NY, USA
{lipishan,fxu3,inwogu}@buffalo.edu

Abstract. Both manual (relating to the use of hands) and non-manual markers (NMM), such as facial expressions or mouthing cues, are important for providing the complete meaning of phrases in American Sign Language (ASL). Efforts have been made in advancing sign language to spoken/written language understanding, but most of these have primarily focused on manual features only. In this work, using advanced neural machine translation methods, we examine and report on the extent to which facial expressions contribute to understanding sign language phrases. We present a sign language translation architecture consisting of two-stream encoders, with one encoder handling the face and the other handling the upper body (with hands). We propose a new parallel cross-attention decoding mechanism that is useful for quantifying the influence of each input modality on the output. The two streams from the encoder are directed simultaneously to different attention stacks in the decoder. Examining the properties of the parallel cross-attention weights allows us to analyze the importance of facial markers compared to body and hand features during a translating task.

Keywords: Parallel Cross-Attention · Facial Expressions · SLT

1 Introduction

Sign Language is the primary means of communication used by Deaf and Hard of Hearing (DHH) individuals, uses multimodal cues to fully express the intended meaning. In the literature, it has been frequently noted that the linguistics of many sign languages have yet to be as well analyzed and studied when compared to their spoken counterparts [31]. This makes exploring automated sign language understanding an important research field where emphasis can be put on analyzing how other visual cues support the sign gestures [2].

Similar to spoken language, sign language has its own syntax, grammar, and semantics. Spoken language uses speech as one of its main vehicles of transmission. On the contrary, sign language combines visual gestures such as hand shapes and movements, body movement, mouthing cues, and facial expressions.

A. Antonacopoulos et al. (Eds.): ICPR 2024, LNCS 15321, pp. 372–386, 2025.
https://doi.org/10.1007/978-3-031-78305-0_24

These components are divided into two major categories: a) manual markers, which include parameters like hand shapes, palm orientation, hand/arm movement, and hand location changes, and b) non-manual markers, such as facial expressions and mouthing cues.

Current works in sign language translation or generation mainly use only the manual markers, oftentimes ignoring the non-manual aspects of the sign or, at best, including them implicitly [4]. While this may be sufficient to accomplish simple comprehension tasks, they lack the ability to capture the rich expressive power[1] of the sign language.

In this work, we are interested in analyzing the importance of using manual and non-manual markers in Sign Language Translation (SLT). While there have been different transformer-based models in the general pattern recognition/machine learning community that model multimodal signals by fusing them and then decoding the fused embedding, such models are unable to disentangle the individual contributions of each input modality to the output result. Fusion happens, but the extent of the contributions becomes opaque due to the construction of such models.

Hence, in this work, we aim to take advantage of how the cross-attention or (encoder-decoder attention) mechanism in transformers highlights the relationship between the input feature tokens and the plausible predicted output tokens. Using such attention we can easily understand which input feature attends to which output token the most at time t. Our proposed model encodes each input channel separately and then fuses them via *cross-attention in the decoder*. This ensures that each feature embedding is used to predict the final output tokens, thus providing insight into the extent of the individual feature contributions. To the best of our knowledge, there are no such multimodal transformer architectures that can be used to explicitly measure the influence of each encoder input on the resulting decoder output.

We evaluate our proposed method by estimating the influence of manual and nonmanual features during Sign Language translation, tested on the RWTH-PHOENIX-Weather2014T (PHOENIX14T) dataset [4] and the real-life American Sign Language dataset (ASLing) [1]. More details on the datasets are given in Section 4.1. We benchmark our method against other translation methods that use PHEONIX14T, thus demonstrating the viability of the proposed method. We also use the real-life ASLing dataset to provide a more qualitative assessment.

In summary, we propose a novel parallel cross-attention decoder transformer-based approach to analyze the contributions of each component, hand-based sign gestures (manual) and facial expressions (non-manual markers), in the Sign Language translation (SLT) task. While significant efforts have been made to improve the performance numbers in SLT tasks, in this work, our focus remains on conducting a comparative analysis of two major sign language components and how they influence the downstream task of translation.
To this end, the contributions of this work can be summarized as follows:

[1] The expressive power of a language refers to the variety and quantity of concepts that can be represented and communicated in that language.

- We design a general, all-purpose, multi-channel cross-attention (encoder-decoder attention) fusion model useful in multimodal pattern analysis for understanding the influence that each input modality contributes to the resulting end task.
- Using the proposed cross-attention fusion technique, we analyze the role of facial expression in the Sign Language translation task.
- Employing two well-established datasets in the field, we present quantitative and qualitative evaluation approaches for a continuous sign translation task. These rely on the attention weights created by the parallel cross-attention model during inference.

2 Related Works

Multi-Modal Sign Language Translation Sign Language Translation (SLT) focuses on interpreting the signals conveyed in signing videos to spoken/written text. Early works [4] introduced an RNN-based pipeline to solve the SLT problem. Many works [4,6,19,33] began using Transformer-based architectures to further improve the performance. Most of these works have used skeleton-based features focusing on the upper body and hands. Camgoz et al. [5] introduced a framework that allowed for separate input channels to process individual sign language components. Other works made similar advances by also incorporating multiple modalities, though not necessarily multiple components of Sign Language [9] - their two modality streams were the (i) raw RGB frames and (ii) corresponding body keypoint sequence extracted from the frames. Chaudhary et al. [8] built an end-to-end 2-way pipeline to use the sign-translated phrases to generate the original signs.

Manual and Non-Manual Markers There is a common misconception that Sign Language is only communicated using hand gestures [23], but in reality, it is much more complex than this. Similar to phonemes in sound, Sign Language phonology underscores the structure of each sign and the way they are organized. Each sign can be broken down into smaller parts made up of the handshape, hand location, hand/arm movement, palm orientation, and the corresponding nonmanual cues [32]. As outlined in Section 1, this work focuses on understanding the manual signals (especially the upper torso with handshapes and their movements) and non-manual signals, specifically facial expressions as related to sign understanding.

Silva et al. [30] introduced a FACS-based facial expression database for Brazilian Sign Language. Mukushev et al. [24] analyzed similar signs to find if non-manual components can differentiate them distinctly. Zheng et al. [35] attempted to improve the performance of SLT by adding facial expressions as input. Koller et al. [15] modeled mouthing shapes in reference to the sign language recognition task.

3 Method

We introduce a fusion model that is useful for understanding the weightage of each input feature representation during the training process when the model is presented with more than one type of input. The proposed model consists of two encoders and a decoder. The architecture is trained for a Sign Language Translation (SLT) task, which is considered a sequence-to-sequence learning problem [4].

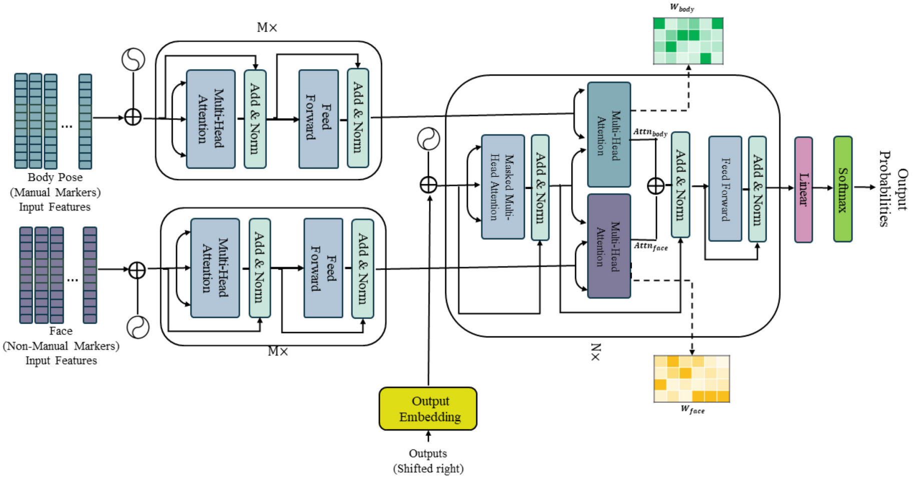


Fig. 1: Overview of the proposed architecture showing the two-stream dual encoder and the dual-cross-attention based decoder.

Figure 1 shows the overview of our proposed architecture, which consists of a) two separate encoders, each for a single feature input, and b) a dual-cross-attention-based decoder for generating spoken language output texts.

3.1 Parallel Cross-Attention Decoder Transformer

The proposed architecture follows the framework of a standard encoder-decoder Transformer but consists of two detached encoders and an added encoder-decoder attention layer. Each encoder is responsible for learning one input feature representation: (i) hands and body pose (manual markers) and (ii) face features (non-manual markers). The model is trained to accomplish the task of SLT by learning the attention weights of individual features using a dual-cross-attention module as the intermediary task. The intuition is that the problem at hand involves understanding how much each of the two separate inputs contributes to the single final task. A similar architecture [16] used a dual-decoder-based transformer for the joint simultaneous tasks of automatic speech recognition (ASR) and speech translation. Note that this significantly differs from our proposed work where we have dual encoders interacting within the cross-attention block.

By simultaneously having two separate attention blocks being responsible for each of the input features, the properties of each attention block can be evaluated to determine how much influence that encoder (hence that modality) has on the final output decoder task.

Two-stream encoding pipeline We make use of two separate encoders to independently learn the representation of each of the input features. Each encoder consists of an input embedding layer followed by a positional embedding and several self-attention and feed-forward network (FFN) layers whose inputs are normalized. Both the encoders follow the same layer configuration as a standard Transformer encoder [33].

The first encoder takes the segments of stacked 3D body landmarks [22] $X_b = ({x_1}^b, {x_2}^b, {x_3}^b, ..., {x_n}^b)$ (further explained in Section 4.1) and the second encoder takes segments of stacked 3D face landmarks [22] $X_f = ({x_1}^f, {x_2}^f, {x_3}^f, ..., {x_n}^f)$ as input sequence. Similar to the standard encoder layers the input sequence is modeled with self-attention and mapped into contextual representation: $z_b = ({z_1}^b, ..., {z_n}^b)$ for the manual markers which are the body and hand joints[2] and $z_f = ({z_1}^f, ..., {z_n}^f)$ for the face landmarks which are the non-manual markers. Each encoder stream is composed of M number of layers.

Decoding pipeline The parallel cross-attention decoder consists of a single decoder with an additional multi-head attention layer. The additional multi-head attention layer paired with the existing multi-head attention layer together are called the **parallel cross-attention layers**. Each parallel attention layer follows a standard encoder-decoder attention layer (also called multi-head attention). An attention layer is defined by a function that maps a query and a pair of key-value to an output vector. The attention function receives the inputs as a key K, a value V, and a query Q. The decoder usually performs two types of attention functions: a) a self-attention is performed on the shifted output sequence, where the inputs K, V, and Q are the same, and b) a cross-attention, also known as encoder-decoder attention, which maps the context representations with the output sequence that are received from the self-attention layer.

In order to understand the importance of any feature against the output sequence, cross-attention values are considered. Taking inspiration from parallel combination strategy [20], we propose a model that will learn the context representations from both features simultaneously, along with the weightage the model puts on each feature when deciding on the final output sequence. To formulate the problem, we consider the context representations from both encoders as separate inputs to their respective attention function.

The attention function is given as:

$$Attn(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{1}$$

[2] Here landmarks and joints can be used interchangeably

In the proposed parallel cross-attention decoder, each attention function is responsible for attending to the output context representation from each encoder. We let $Attn_{body}$ be the attention layer that takes the output context representation z_b of the body encoder as the key-value pair input. Similarly, $Attn_{face}$ is the attention layer that attends to the output context representation z_f from the face encoder as the key-value pair. In both cases, the query value is received by the normalized masked self-attention output from the previous shifted token of the sequence.

To formulate the concept of parallel cross-attention decoder, we denote the inputs as follows:

$Q_{shifted}$ is the query value received from the shifted output sequence
$K_{(z_f)}, V_{(z_f)}$ is the key-value pair received from the face encoder stream
$K_{(z_b)}, V_{(z_b)}$ is the key-value pair received from the body encoder stream
Thus, each of the attention functions in the parallel cross-attention decoder is:

$$Attn_{body} = softmax\left(\frac{Q_{shifted}K^T_{(z_b)}}{\sqrt{d_k}}\right)V(z_b) \tag{2}$$

$$Attn_{face} = softmax\left(\frac{Q_{shifted}K^T_{(z_f)}}{\sqrt{d_k}}\right)V(z_f) \tag{3}$$

The final output from the parallel cross-attention is received by merging both the attention outputs from 2 and 3. We denote the final merged output as $Attention_{final}$. The concatenation operator was used to merge the outputs to get to the final output:

$$Attn_{final} = linear([Attn_{body}; Attn_{face}]) \tag{4}$$

The combined attention output is normalized and then passed onto the feed-forward network to predict the next token auto-regressively. We simultaneously look at the attention weights from both the attention functions for the body and face. For this, we focus on the correlation computed between the query $(Q_{shifted})$ and each key value - K_{body} & K_{face}. Specifically, we observe the outputs of the Scaled Dot-Product Attention [33] after the softmax layer. We formulate the attention weights as follows:

$$w_{body} = softmax\left(\frac{Q_{shifted}K^T_{(z_b)}}{\sqrt{d_k}}\right) \tag{5}$$

$$w_{face} = softmax\left(\frac{Q_{shifted}K^T_{(z_f)}}{\sqrt{d_k}}\right) \tag{6}$$

We use the weights from Equations 5 and 6 to analyze the role of manual and non-manual markers in understanding sign language and present our findings in Section 5.

4 Implementation and Evaluation Details

4.1 Datasets

First, we use the benchmark dataset, PHOENIX14**T** [4] to evaluate and establish credibility for our proposed translation model. We train our proposed parallel cross-attention decoder transformer with 7096 training, 519 validation, and 642 test samples. The samples were collected from the weather forecast airings and performed by nine different signers in German Sign Language (GSL)[3]. Along with the samples, the dataset contains their corresponding German translations and gloss annotations.

Next, we evaluate the proposed model on the more real-life, unconstrained American Sign Language dataset, ASLing [1]. This dataset consists of 1027 training samples and 257 testing samples. The samples were performed by 7 native signers and collected at 10 frames per second.

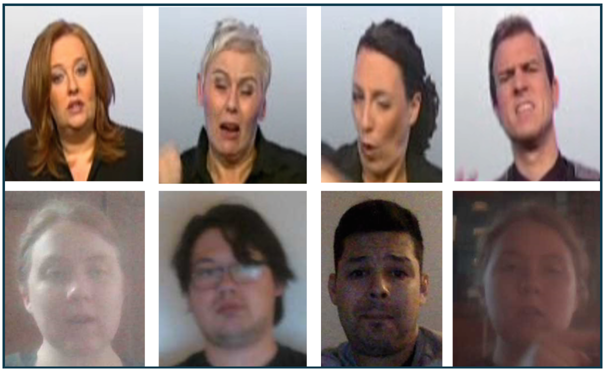


Fig. 2: Face crop samples from the datasets. Top: from the Phoenix2014 dataset; Bottom: from the ASLing dataset

4.2 2-stage Input Feature Extraction Process

3D keypoints To accurately track the motion of each of the markers (body and face), we make use of 3D joint features $j = (x, y, z)$. We use [22] to extract both body and face 3D joints for the input sequence X_M. To independently understand the weightage of each of the modalities, we separate the body and face joints for individual video input sequences. Given the 3D graphical representations of the input features, we employ a Spatial-Temporal Graphical Convolution Network (STGCN) [34]. Indices are selected based on points depicted in Figure 3.

[3] German Sign Language is also known as Deutsche Gebärdensprache (DGS).

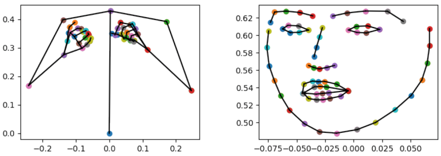


Fig. 3: The (x, y) joint plots for body (left) and face (right). Note that we use 3D points (x, y, z), in our analysis.

For the manual markers, we consider only the upper torso with hands. The body keypoints follow [29] format; we eliminate the foot, ankle, and knee keypoints to get a total of 48 body keypoints including hands. We consider all 72 face keypoints for the non-manual markers to create the skeleton. There is a total of 120 selected keypoints which are used as the input to the embedding extractor. To build a connection between the body and face keypoints, we build a custom tree structure which is used in the processing of a Spatio-Temporal Graphical Convolution Network (STGCN).

We pre-train the STGCN on a word-level Sign Language recognition dataset (WLASL) [18] in order to learn general sign language motion representations. The resultant vector per temporal window is of length 1024. Figure 4 shows the process of feature extraction. The right side of the process is pretrained twice; once on manual input features and the second on non-manual input features, to get 1024 feature embeddings per modality.

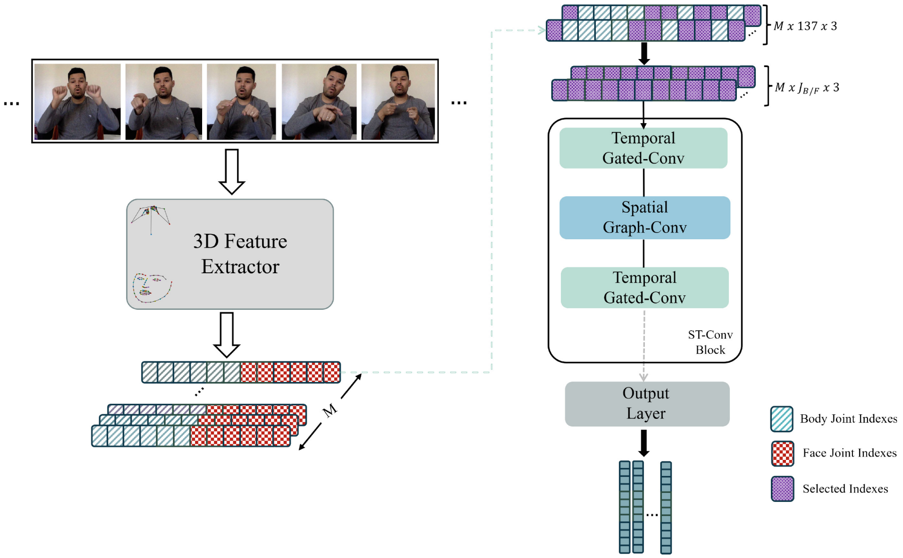


Fig. 4: Overview of the 2-stage input feature extraction process.

Rotation Matrix Another feature we consider with the human skeleton structure is the 3D rotation matrix, a special orthogonal 3×3 matrix ($SO(3)$), that represents the rotation of one joint in Euclidean space. A chain of rotation matrices from the origin joint (pelvis) on the skeleton is used to represent the human pose for each video frame.

Recent works [17] [3] have shown the effectiveness of different rotation representations of $SO(3)$ for gradient-based optimization in neural network learning. Specifically, a rotation representation that contains less than five dimensions suffers from discontinuities in the real Euclidean space, which leads the gradient of the loss function to blow up [11]. The 3×3 rotation matrix we considered in this work contains 9 parameters, which bakes in the redundancies needed to deal with the discontinuous issue as compared to other rotation representations.

To make a fair comparison with 3D keypoint features, we follow the same protocol to consider only the upper body and hand joints. We also pre-train the STGCN with the WLASL dataset to learn the best motion representations from the rotation matrix.

Width In continuous sign language videos, individual signs span over multiple frames; hence, for analyzing the representation of each sign, we exploit its spatiotemporal properties by modeling it with a multi-scale STGCN. We use the kernel window in the Temporal Gated Convolution layer of the STGCN to obtain these segments. We consider a given video of length N, channels or dimension size c, and the number of 3d joints v. We define a 1D convolution kernel τ which maps the input chunks to a single output Y [34].

4.3 Training and Network Details

Our proposed network is trained using the Pytorch framework [28] and follows the FAIRSEQ transformer setup [25]. We use Adam optimizer [14] with a batch size of 32 and initialize the learning rate to 10^{-4} with a weight decay of 10^{-4}. The number of decoder attention heads was set to 12 and the best head was selected to analyze the output attention weights of each modality. The model was trained on a single NVIDIA RTX 3090 Ti processor.

4.4 Evaluation Metrics

We use the two popular language evaluation metrics to access our performance: BLEU (BiLingual Evaluation Understudy) [26] and ROUGE-L [21]. BLEU computes a modified n-gram precision where for each candidate n-gram a maximum corresponding reference is counted. ROUGE-L was usd to measure the sentence-wise similarity based on the longest common sub-sequence statistics between a candidate translation and a set of reference translations.

5 Experiments and Results

The proposed model is evaluated on the Phoenix2014T and ASLing datasets, and both quantitative and qualitative (explainable) results are reported. We experiment with different settings of input modalities on the model, reporting results on manual markers only, non-manual markers only, and manual + non-manual markers. The attention weights for each modality is dissected and our findings are presented via the extracted attention plots

5.1 Quantitative

We test the proposed model with benchmark Phoenix2014T dataset and achieve a BLEU4 [27] score of 11.27 on the task of Sign Language translation; the complete results are shown in Table Table 1. Table 2 shows the performance scores on the ASLing dataset. Due to its noisy nature and smaller size, we fine-tune the model on ASLing by transfer learning from Phoenix2014T. Recall that the aim of this work is to understand the structure and roles of different modalities in Sign Language Understanding.

Table 1: Performance of proposed method on Phoenix2014T dataset

Method	**Width**	**Modality**	**Feature**	**Bleu-1**	**Bleu-2**	**Bleu-3**	**Bleu-4**	**Rouge-L**
Sign2Text [4]	-	Manual	-	32.24	19.03	12.83	9.58	31.80
MCT [5]	-	Multi-channel	-	-	-	-	18.51	43.57
Skeletor [13]	-	Manual	2d to 3d lifting	31.86	19.11	13.49	10.35	31.80
TSPNet [19] (Single)	{8}	-	RGB [7]	30.29	17.75	12.35	9.41	28.93
	{9}	-	RGB [7]	23.87	15.49	11.08	8.71	24.7
	{12}	-	RGB [7]	29.02	17.03	12.08	9.39	28.10
	{16}	-	RGB [7]	35.52	20.33	14.75	11.61	32.36
Parallel Cross-Attention Decoder (Ours)	{9}	Manual	3dkeypoints	29.40	17.23	11.98	9.82	23.4
		Non-manual	3dkeypoints	22.35	11.32	7.68	6.25	15.5
		Manual + Non-manual	3dkeypoints	29.81	17.15	11.65	9.87	-
		Manual	rotmat9d	24.65	13.24	9.07	7.44	18.4
		Manual + Non-manual	3d-keypoints +rotmat9d	32.79	19.91	13.7	11.27	-

The intuition behind using the Phoenix2014T dataset is to verify the model's end performance and guarantee that the attention weights are learned correctly. We experimented with different temporal width settings with STGCN, and based on the accuracy achieved, we selected the width to be used in the model. We also experimented with different feature extraction methods, i.e., 3d-keypoints and 3d rotation matrix (rotmat9d). The reported numbers for non-manual markers for the ASLing dataset were substandard due to its noisy nature. Although there were several images with clear, full-frontal face views, the number of face images with poor lighting or pose conditions dominated. Nevertheless, we were still able to identify patterns consistent with the Sign Language semantics. We believe that the frames' quality can directly impact the feature quality; hence, the model attends poorly when the image quality is low, as shown in Figure 2.

Table 2: Performance of proposed method on ASLing Transfer learned on Phoenix2014T

Method	Modality	Feature	Bleu-1	Bleu-2	Bleu-3	Bleu-4
CFDF	Manual	2d-keypoint + Optical Flow + ResNet	22.39	15.96	13.56	12.25
Parallel Cross-Attention Decoder (Ours)	Manual +Non-manual	rotmat9d	21.82	16.08	13.67	12.30

5.2 Qualitative

We observe the behavior of our proposed model by plotting the learned attention weights for each modality. The correlation is strong between a sequence frame and the decoder output token when that frame consists of strong facial feature movements that influence the output task. This indicates that along with body motions, often when signing, facial features can also contribute significantly to the output task (SLT in this case).

Figure 5 shows the attention weights for one sample from the testing set of the Phoenix2014T dataset. The x-axis of the plot represents the input sequence frames, and the y-axis represents the decoder output tokens. The darker blocks in Figure 5 (bottom) demonstrate that the non-manual input features (facial expressions) are attending more in the correlations to the output tokens.

6 Discussion and Conclusion

Fusion Techniques Most of the previous works using multi-modalities have used early or late fusion of features. In the early fusion method, the features from different modalities are fused to create a single representation of the sequence. This implicitly helps the translation model build meaningful contextual relationships between the decoder output token and fused features. Although this model works well for fusing multimodal inputs, it cannot determine the extent of influence of each of the input modalities. Camgoz et al.[5] noted that performing this type of late fusion does not always yield better results. Contrary to this, our proposed model performs fusion at decoding time, not to learn about feature representations but to extract learned attention weights for each modality.

We also show qualitative results on ASLing dataset in Figure 6. The plots show the decoder output tokens plotted against the sequence frames. These plots show the attention weights for the phrase: ***"Christmas is my favorite season!"***. For the duration of the input sequence in this example, based on the

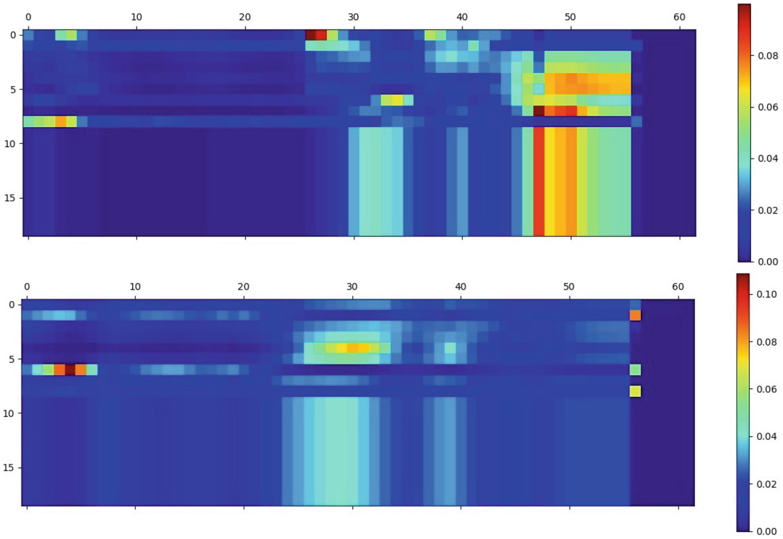


Fig. 5: Learned attention weights for Phoenix2014T dataset. Top: attention weights based on the input manual markers (body features); bottom: attention weights based on the input non-manual markers (facial features)

attention weights, we see that the facial features attend to the output token a little more than the body features. This behavior can be seen when the signer asks a question or conveys excitement or enthusiasm. A larger attention weight value for a particular modality at time t indicates that this feature contributed more to constructing the context representation between encoder and decoder outputs at time t.

Performance Metrics Recent works have achieved better results than the proposed method on Phoenix-2014T, have used gloss annotations to supervise their training [12], [10]. Needless to say, it is assumed that the multi-modality methods are expected to give a better performance because of the added modality. However, the performance of adding a modality can not always be guaranteed. Many factors contribute to this; in the case of ASLing, the dataset was collected in the wild (real-life setting) and is noisy Figure 2 (Bottom). On the contrary, the Phoenix2014T dataset was collected in a more constrained environment where the participants wore dark clothing to contrast with the background; the environment also had controlled lighting - See Figure 2 (Top). This can pose a challenge when analyzing the ASLing dataset. Additionally, non-manual markers convey more than just facial expressions they also construct the grammatical meaning of a sign. This can be a complex structure to decode and understand if the rules of a language are not learned.

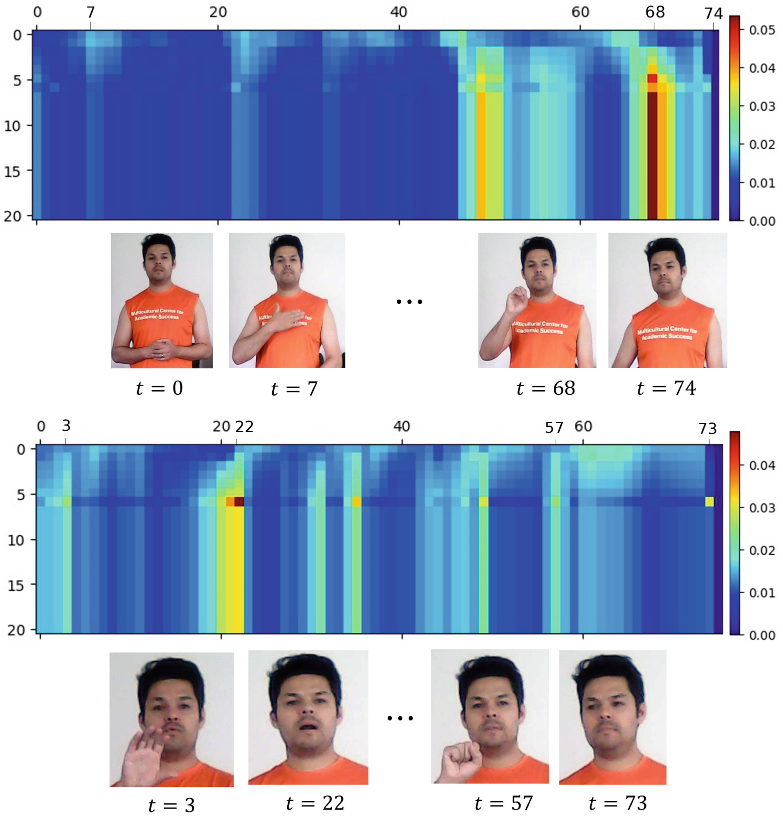


Fig. 6: Learned attention weights for ASLing dataset. Top: attention weights based on the input manual markers (body features); bottom: attention weights based on the input non-manual markers (facial features).

Conclusion

In this paper, in an attempt to develop an influence model in a multimodal architecture, we introduced a dual encoder model, along with a parallel cross-attention decoder, to study the contributions of manual and non-manual features in Sign Language translation. Through the parallel cross-attention mechanism, we are able to retrieve the attention weights for individual modalities, and while the underlying goal is Sign Language translation (proven via quantitative measures), the proposed parallel cross-attention mechanism proved exceptionally useful in estimating the contribution of influence that each modality had on the decoder output during inference. This allowed us to measure the influence of facial expressions on sign translation for different types of signed input phrases. This attribute of the model is its major distinguishing factor among other existing Transformer-based architectures.

References

1. Ananthanarayana, T., Kotecha, N., Srivastava, P., Chaudhary, L., Wilkins, N., Nwogu, I.: Dynamic cross-feature fusion for american sign language translation. In: FG 2021. IEEE (2021)
2. Bragg, D., Koller, O., Bellard, M., Berke, L., Boudreault, P., Braffort, A., Caselli, N.K., Huenerfauth, M., Kacorri, H., Verhoef, T., Vogler, C., Morris, M.R.: Sign language recognition, generation, and translation: An interdisciplinary perspective. ACM SIGACCESS Conference on Computers and Accessibility (2019)
3. Brunton, S.L., Budišić, M., Kaiser, E., Kutz, J.N.: Modern koopman theory for dynamical systems. arXiv preprint arXiv:2102.12086 (2021)
4. Camgoz, N.C., Hadfield, S., Koller, O., Ney, H., Bowden, R.: Neural sign language translation. In: 2018 CVPR. pp. 7784–7793 (2018)
5. Camgoz, N.C., Koller, O., Hadfield, S., Bowden, R.: Multi-channel transformers for multi-articulatory sign language translation. In: Bartoli, A., Fusiello, A. (eds.) ECCV Workshops (2020)
6. Camgoz, N.C., Koller, O., Hadfield, S., Bowden, R.: Sign language transformers: Joint end-to-end sign language recognition and translation. In: CVPR (2020)
7. Carreira, J., Zisserman, A.: Quo vadis, action recognition? a new model and the kinetics dataset. In: CVPR (2017)
8. Chaudhary, L., Ananthanarayana, T., Hoq, E., Nwogu, I.: Signnet ii: A transformer-based two-way sign language translation model. IEEE Trans. Pattern Anal. Mach. Intell. **45**(11), 12896–12907 (2023). https://doi.org/10.1109/TPAMI.2022.3232389
9. Chen, Y., Zuo, R., Wei, F., Wu, Y., LIU, S., Mak, B.: Two-stream network for sign language recognition and translation. In: Oh, A.H., Agarwal, A., Belgrave, D., Cho, K. (eds.) NeurIPS (2022)
10. Chen, Y., Zuo, R., Wei, F., Wu, Y., Liu, S., Mak, B.: Two-stream network for sign language recognition and translation. NeurIPS (2022)
11. Geist, A.R., Frey, J., Zobro, M., Levina, A., Martius, G.: Learning with 3d rotations, a hitchhiker's guide to so (3). arXiv preprint arXiv:2404.11735 (2024)
12. Guan, M., Wang, Y., Ma, G., Liu, J., Sun, M.: Multi-stream keypoint attention network for sign language recognition and translation (2024)
13. Jiang, T., Camgöz, N.C., Bowden, R.: Skeletor: Skeletal transformers for robust body-pose estimation. In: 2021 CVPRW. pp. 3389–3397 (2021)
14. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. In: Bengio, Y., LeCun, Y. (eds.) 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings (2015)
15. Koller, O., Ney, H., Bowden, R.: Deep learning of mouth shapes for sign language. In: 2015 IEEE International Conference on Computer Vision Workshop (ICCVW) (2015)
16. Le, H., Pino, J.M., Wang, C., Gu, J., Schwab, D., Besacier, L.: Dual-decoder transformer for joint automatic speech recognition and multilingual speech translation. In: International Conference on Computational Linguistics (2020)
17. Levinson, J., Esteves, C., Chen, K., Snavely, N., Kanazawa, A., Rostamizadeh, A., Makadia, A.: An analysis of svd for deep rotation estimation. NeurIPS **33** (2020)
18. Li, D., Rodriguez, C., Yu, X., Li, H.: Word-level deep sign language recognition from video: A new large-scale dataset and methods comparison. In: WACV (2020)
19. Li, D., Xu, C., Yu, X., Zhang, K., Swift, B., Suominen, H., Li, H.: Tspnet: Hierarchical feature learning via temporal semantic pyramid for sign language translation. In: Advances in Neural Information Processing Systems. vol. 33 (2020)

20. Libovický, J., Helcl, J., Mareek, D.: Input combination strategies for multi-source transformer decoder. In: Conference on Machine Translation (2018)
21. Lin, C.Y., Och, F.J.: Automatic evaluation of machine translation quality using longest common subsequence and skip-bigram statistics. In: Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL-04). pp. 605–612. Barcelona, Spain (Jul 2004)
22. Lin, J., Zeng, A., Wang, H., Zhang, L., Li, Y.: One-stage 3d whole-body mesh recovery with component aware transformer. In: CVPR. pp. 21159–21168 (2023)
23. Mohr, S.: Chapter 3. Non-manuals in Sign Languages – Theoretical Background, pp. 31–63. De Gruyter Mouton, Berlin, Boston (2014)
24. Mukushev, M., Sabyrov, A., Imashev, A., Koishybay, K., Kimmelman, V., Sandygulova, A.: Evaluation of manual and non-manual components for sign language recognition
25. Ott, M., Edunov, S., Baevski, A., Fan, A., Gross, S., Ng, N., Grangier, D., Auli, M.: fairseq: A fast, extensible toolkit for sequence modeling. In: Ammar, W., Louis, A., Mostafazadeh, N. (eds.) Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations). pp. 48–53. Association for Computational Linguistics (Jun 2019)
26. Papineni, K., Roukos, S., Ward, T., Zhu, W.J.: Bleu: a method for automatic evaluation of machine translation. p. 311–318. ACL '02, ACL, USA (2002)
27. Papineni, K., Roukos, S., Ward, T., Zhu, W.J.: Bleu: a method for automatic evaluation of machine translation. In: Isabelle, P., Charniak, E., Lin, D. (eds.) ACL, pp. 311–318. Association for Computational Linguistics, Philadelphia, Pennsylvania, USA (Jul (2002)
28. Paszke, A., Gross, S., Chintala, S., Chanan, G., Yang, E., DeVito, Z., Lin, Z., Desmaison, A., Antiga, L., Lerer, A.: Automatic differentiation in pytorch. In: NeurIPS 2017 Workshop on Autodiff (2017)
29. Pavlakos, G., Choutas, V., Ghorbani, N., Bolkart, T., Osman, A.A.A., Tzionas, D., Black, M.J.: Expressive body capture: 3D hands, face, and body from a single image. In: CVPR. pp. 10975–10985 (2019)
30. Silva, E.P.d., Costa, P.D.P., Kumada, K.M.O., De Martino, J.M.: Silfa: Sign language facial action database for the development of assistive technologies for the deaf. In: FG 2020 (2020)
31. Stokoe, W.C.: Sign language structure. Annual Review of Anthropology **9** (1980)
32. VALLI, C., LUCAS, C., MULROONEY, K.J., VILLANUEVA, M.: Linguistics of American Sign Language, 5th Ed.: An Introduction. Gallaudet University Press (2011)
33. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L.u., Polosukhin, I.: Attention is all you need. In: Guyon, I., Luxburg, U.V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R. (eds.) NeurIPS. vol. 30 (2017)
34. Yu, B., Yin, H., Zhu, Z.: Spatio-temporal graph convolutional networks: A deep learning framework for traffic forecasting. In: Proceedings of the 27th International Joint Conference on Artificial Intelligence (IJCAI) (2018)
35. Zheng, J., Chen, Y., Wu, C., Shi, X., Kamal, S.M.: Enhancing neural sign language translation by highlighting the facial expression information. Neurocomputing **464**, 462–472 (2021)

GolfPose: From Regular Posture to Golf Swing Posture

Ming-Han Lee(✉), Yu-Chen Zhang, Kun-Ru Wu, and Yu-Chee Tseng

Department of Computer Science, National Yang Ming Chiao Tung University,
No. 1001 University Road, Hsinchu, Taiwan
{mhlee.cs09,yuchen2856.cs10,wufish,yctseng}@nycu.edu.tw

Abstract. While there already exist a number of 2D and 3D pose estimation models with high accuracy, in special domains like sports, which usually require even higher accuracy, there are still spaces to be improved. Existing pose models primarily focus on regular daily activities, which, when being applied to precision sports, such as golf swings, still face limitations. In fact, the rare poses and self-occlusions in golf swing videos can easily mislead regular pose models. To overcome these challenges, we develop a small (2D and 3D) GolfSwing dataset that includes both golfer and club poses. We then fine-tune state-of-the-art 2D and 3D posture models, including HRNet, ViTPose, DEKR, and MixSTE, by GolfSwing into a set of models called GolfPose for golfer-club pose estimation with much higher accuracy. Such a simple-yet-effective method may be generalized to other sports with self-occluded properties. Code is available at https://github.com/MingHanLee/GolfPose.

Keywords: Human Pose Estimation · Golf · Motion Capture · Precision Sports · Self Occlusion

1 Introduction

Human pose estimation (HPE) has been intensively studied in computer vision and sensor fields. Solutions can be categorized as 2D and 3D ones. Image-based 2D HPE models are proposed in[2,8,31], while 3D postures can be derived by regression [12,17,32] or by 2D-to-3D lifting [25,28]. There are wide ranges of pose applications in sports, including using rugby players' poses to evaluate the risk of concussion during a tackle [27], predicting 3D flight trajectory of badminton [20], incorporating a 3D geometry of the scene to enhance the accuracy of 3D HPE [1], and comparing the pose differences between professional and amateur runners using PoseCoach [19].

In this work, we consider the inference of golf swing videos taken by an off-the-shelf RGB camera. Golf has been increasingly popular in recent years. Golf swings directly impact performance. The studies [26,44] utilized motion capture systems to collect golf swing poses and analyze its relation with injuries. References [13,14] used HPE to identify key frames in golf swing videos and assess

A. Antonacopoulos et al. (Eds.): ICPR 2024, LNCS 15321, pp. 387–402, 2025.
https://doi.org/10.1007/978-3-031-78305-0_25

the effectiveness of a swing. How to employ deep learning to coach a beginner's swings based on experts' ones is addressed in [15]. A similar study based on motion capturing is in [16]. The GolfDB dataset [23] consists of 1,400 videos of professional golfers' swings with event frames and bounding boxes labeled. However, the dataset is 2-dimensional and lacks pose keypoint annotations.

Our goal is to estimate both 2D and 3D golfer-with-club postures through a normal RGB camera. To the best of our knowledge, there is no dataset containing all such annotations. We first develop a small *GolfSwing* dataset by a high-quality motion capture system, which features ground truth of 3D golfer-with-club keypoints. There are 17 keypoints for golfer and 5 keypoints for club. We further synchronize these information with normal RGB cameras and project these 3D keypoints to 2D ones as the ground truth. *GolfSwing* enables us to derive a set of more accurate 2D and 3D models, called *GolfPose*, to infer golfer-with-club keypoints through regular videos. In particular, we take a fine-tuning approach. First, a number of 2D state-of-the-art HPE models are fine-tuned, including HRNet [31], ViTPose-H [37], and DEKR [8]. Second, we include club keypoints and fine-tune MixSTE [39], the state-of-the-art 2D-3D lifting model. The results may facilitate various downstream golf applications.

We test these 2D and 3D models fine-tuned from *GolfSwing*. Our experimental results indicate that the original 3D MPJPE of MixSTE can be reduced from 109.4 mm to 35.6 mm and, if we further include club with golfer, the 3D MPJPE can be reduced to a 32.3 mm. For the 2D case, the original mAP of the tested 2D models can be increased from the range of 0.669-0.706 to 0.877-0.936; if we further include club keypoints, the mAP can be increased to 0.918-0.956. This simple-yet-effective approach not only validates the value of *GolfSwing*, but also indicates the feasibility of pretraining a 2D/3D pose model on large datasets like Human3.6M [11] and TotalCapture [33], which primarily focus on regular daily poses, followed by fine-tuning it with a small human-with-object dataset. The results can also be generalized to other precision sports that suffer serious self-occlusion effects, like tennis, badminton, and cricket.

2 Related Work

Motion Capture Systems. They can be categorized as marker-based, markerless, and inertial sensor-based. Marker-based systems [34] utilize reflective materials to facilitate tracking. Through multiple cameras, the 3D locations of markers are positioned by triangulation. While accurate, such systems are more costly and difficult to set up. Markerless systems [3] do not require markers and track the optical flows of pixels in 2D image spaces for constructing 3D positions. Inertial sensor-based solutions are less costly and provide more degrees of freedom [30]. However, error accumulation is a persistent problem.

Our *GolfSwing* dataset was recorded concurrently by RGB cameras and Vicon cameras (a marker-based system), thus featuring both 2D and 3D ground truth. Markers are attached to both golfer and club. We follow the configurations in [11] in our setup.

Golf Kinematics. A lot of studies tried to understand golf kinematics. To study golf swings and injuries, [44] recorded LPGA and PGA golfers' motions and collected statistics including angles and angular velocities of swings. The differences in injury risks and swing techniques among male and female professional golfers on injury regions were studied. The correlation between lumbar and hip joint rotation during a swing and its association with lower back pain was investigated in [26]. To help beginners to correct their poses, HRNet with Simplebaseline3d was employed in [15] to infer 3D poses in GolfDB [23]. Through [16], a learner's poses can be synchronized with coaches' in database, thus providing visualization assistance to learners.

2D HPE. 2D HPE can be broadly categorized into two approaches: *top-down* and *bottom-up*. The top-down approach [31,35,37] consists of two stages: object detection and pose estimation. It transforms multi-person pose estimation into single-person estimation. It typically achieves higher accuracy but incurs higher computing cost. The bottom-up approach [2,8,38] first estimates all keypoints, followed by poses construction. This approach is faster, but generally less accurate.

3D HPE. Monocular 3D HPE has been widely explored. Solutions can be categorized as one- and two-stage ones. The one-stage approaches [12,17] directly regress 3D skeletons from input without intermediate 2D skeleton representations and are thus more computing-intensive. Two-stage approaches first employ a 2D pose detector to identify skeletons and then elevate 2D skeleton sequences to 3D ones. References [36,41] try to predict 3D skeletons directly from 2D skeletons, and are thus highly sensitive to 2D detection accuracy. Since temporal information of continuous skeletons may reduce depth ambiguity, TCN [28] conducts dilated convolutions on adjacent 2D skeletons to estimate 3D ones. PoseFormer [43] proposes a spatial-temporal transformer encoder to capture skeleton structure and temporal activity. Also based on transformer, MixSTE[39] focuses on the temporal features of individual keypoints and spatial features in each 2D skeleton. Following the recent trend, our GolfPose takes a two-stage approach.

Human Pose Datasets. Consisting of 200,000 images and 250,000 person instances, COCO Keypoints [18] defines 17 2D human keypoints and includes annotations for occlusion situations, enabling significant progress under challenging conditions. For 3D datasets, Human3.6M [11] contains large indoor scenarios with 15 daily actions. Also with 17 human keypoints, it includes various data types such as RGB images, human silhouette, bounding box, depth, 3D pose, and 3D laser-scanned human models. MPI-INF-3DHP [24] incorporates both indoor and outdoor scenes with diverse human poses, clothing, and occlusions. TotalCapture [33] provides human keypoints, activity types, and synchronized sensor data. 3DPW [22] is a 3D dataset collected from handheld cameras with sensors attached to human limbs in outdoor environments. SportsPose [10] consists of five types of sports in dynamic scenes.

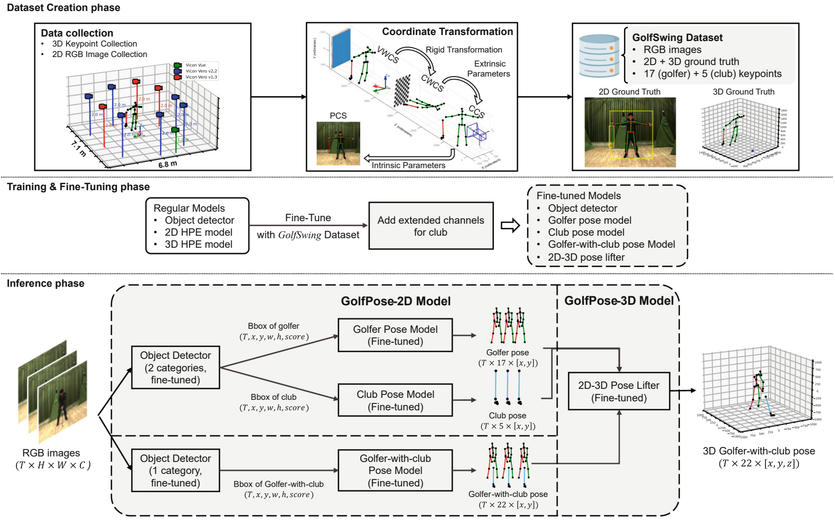


Fig. 1. Framework of *GolfPose*. (Blocks marked by gray represent our contributions.)

3 Methodology

Our goal is not only to enhance the performance of existing 2D and 3D HPE models but also to include keypoints of club. Fig. 1 shows our research framework. There are three phases. The first phase is to derive *GolfSwing* by Vicon cameras [34] plus regular RGB cameras. In the second phase, we will fine-tune detectors and pose models. The third phase is, from normal RGB videos, to infer golfer-with-club keypoints and, for the case of 3D, to conduct 2D-to-3D lifting.

3.1 GolfSwing Dataset

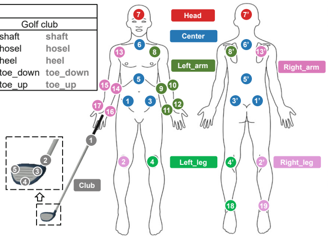
GolfSwing is collected concurrently by 9 Vicon infrared cameras and 2 RGB cameras that are time-synchronized. The environment setup and equipment specifications are shown in Fig. 1. The infrared cameras are placed around the golfer, while the RGB cameras are placed in the front and the side of the golfer. The area size is about 6.8m x 7.1m. The golfer stands within the capture region, utilizing a 7-iron club.

Before recording, we calibrate all Vicon cameras with a Vicon wand. A center point on the ground is regarded as the 3D origin. There are 6 volunteer students serving as golfer. Each volunteer is tagged by 28 markers for 3D trajectory tracking. The marker placement is designed similar to Human3.6M, from which we can calculate 17 keypoints as ground truth. In addition, club is tagged by 5 markers for keypoint tracking. The details are depicted in Fig. 2.

Post-processing is required because a marker has to be captured by at least two Vicon cameras in order to reconstruct its 3D location. Due to the speciality of golf sports, markers can be easily occluded during a swing. Missing markers

GolfSwing Dataset						
	Training Set				Test Set	
Subjects	Subject 1 (S1)	Subject 2 (S2)	Subject 3 (S3)	Subject 4 (S4)	Subject 5 (S5)	Subject 6 (S6)
Trials	5	3	4	3	3	2
Frames	3892	1560	4794	3536	3195	761
Total	15 trails, 13782 frames				5 trails, 3956 frames	

GolfSwing Keypoints						
Person				Golf club		
p1	hip	r_hip	l_hip	c1	shaft	shaft
p2	r_hip	center1	center1'	c2	hosel	hosel
p3	r_knee	right_leg2	right_leg2'	c3	heel	heel
p4	r_ankle	right_leg19		c4	toe_down	toe_down
p5	l_hip	center3	center3'	c5	toe_up	toe_up
p6	l_knee	left_leg4	left_leg4'			
p7	l_ankle	left_leg18				
p8	spine	center5	center5'			
p9	neck	center6	center6'			
p10	neck_base	head7	center6			
p11	head	head7	head7'			
p12	l_shoulder	left_arm8	left_arm8'			
p13	l_elbow	left_arm9	left_arm10			
p14	l_wrist	left_arm11	left_arm12			
p15	r_shoulder	right_arm13	right_arm13'			
p16	r_elbow	right_arm14	right_arm15			
p17	r_wrist	right_arm16	right_arm17			



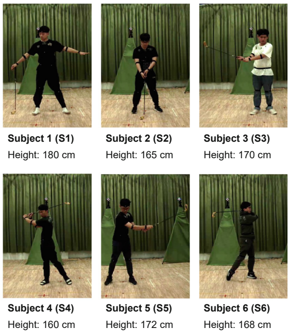


Fig. 2. The specifications of *GolfSwing*.

are replaced using Vicon Nexus's algorithms. In the end, we obtain a set of highly accurate 3D golf swing keypoints as ground truth.

The above steps have led to 3D keypoint ground truth. The last step is to perform coordinate transformation to produce 2D keypoint ground truth. This is done by projecting 3D keypoints onto RGB images. We follow Zhang's calibration algorithm [40] and define four coordinate systems (Fig. 1):

1. *Vicon World Coordinate System (VWCS)*: the 3D coordinate system of Vicon cameras, with the calibration wand as the origin.
2. *Checkerboard World Coordinate System (CWCS)*: the 3D coordinate system to relate real world with RGB cameras via an external checkerboard.
3. *Camera Coordinate System (CCS)*: the 3D coordinate system used by RGB cameras.
4. *Pixel Coordinate System (PCS)*: the 2D coordinate system of RGB images, with the top-left corner as the origin.

Then we conduct three coordinate transformations. The first one is VWCS-to-CWCS transformation. We place Vicon markers at the origin, x-axis, and y-axis of *CWCS* as the transformation basis for *VWCS*. By these markers, we calculate the rotation matrix $R' \in \mathbb{R}^{3\times3}$ and translation matrix $T' \in \mathbb{R}^{3\times1}$, which lead to the Rigid Transformation Matrix $C \in \mathbb{R}^{4\times4}$:

$$C = \begin{bmatrix} R'_{3\times3} & T'_{3\times1} \\ 0_{1\times3} & 0_{1\times1} \end{bmatrix} \tag{1}$$

The second one is CWCS-to-CCS conversion. The Extrinsic Matrix $E \in \mathbb{R}^{4\times4}$ is employed [40]. It is also composed of a rotation matrix $R \in \mathbb{R}^{3\times3}$ and a translation matrix $T \in \mathbb{R}^{3\times1}$, and can be denoted by:

$$E = \begin{bmatrix} R_{3\times3} & T_{3\times1} \\ 0_{1\times3} & 0_{1\times1} \end{bmatrix} \tag{2}$$

The third one is CCS-to-PCS transformation. We utilize the Intrinsic Matrix $K \in \mathbb{R}^{3\times4}$, which consists of the focal length (f_x, f_y) and principal point (c_x, c_y).

It is computed during the calibration algorithm [40]:

$$K = \begin{bmatrix} f_x & 0 & c_x & 0 \\ 0 & f_y & c_y & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix} \tag{3}$$

By combining the above transformation matrices, we derive the projection from a 3D point onto the 2D PCS:

$$s \begin{bmatrix} u \\ v \\ 1 \end{bmatrix} = KEC \begin{bmatrix} x_w \\ y_w \\ z_w \\ 1 \end{bmatrix}, \tag{4}$$

where $\begin{bmatrix} x_w \; y_w \; z_w \; 1 \end{bmatrix}^T$ is a 3D point in *VWCS*, $\begin{bmatrix} u \; v \; 1 \end{bmatrix}^T$ is its corresponding 2D point in PCS, and s is the scale factor referring to the ratio of the physical measurement unit to the image unit. By projecting all 3D keypoints to PCS, we obtain 2D keypoint ground truth of *GolfSwing*. From the above 2D keypoints, we further calculate the bounding boxes of golfer and club as ground truth.

Overall, *GolfSwing* comprises 6 golfers of heights 160-180 cm, who had taken 2-4 sports classes or joined school sports teams. Each subject swung 7 times, yielding a total of 42 trails. We asked volunteers to swing differently each time. After manual curation, 20 highly accurate trails were collected. We followed “cross-subject” split, with 4 for training and 2 for testing. 17,738 frames were collected, with 13,782 (78%) for training and 3,956 (22%) for testing.

To summarize, there are several challenges during data collection: (i) players’ diversity, (ii) recording environments, (iii) fast-moving swings, (iv) missing keypoints in Vicon videos, and (v) opt-in permission required for each volunteer. These are conquered by asking players to swing different each time, cleaning blurry frames (especially for club), and manually making up missing keypoints. Unfortunately, recording in the wild is not feasible currently for Vicon’s IR cameras.

3.2 Model Fine-Tuning

We take a top-down approach [42] for golfer-with-club keypoint detection. With *GolfSwing*, we fine-tune object detector, 2D pose, and 2D-3D lifting models. In particular, there are two alternatives for object and 2D pose detection, one by detecting golfer and club separately and the other by detecting them jointly.

For object detection, we employ Faster R-CNN [29] and YOLOX [7]. Both models are pretrained on the COCO 2017 dataset, which includes 80 object categories. We fine-tune them by *GolfSwing* (2D) dataset. For the separated method, two categories, namely golfer and club, are detected. For the joint method, only one golfer-with-club category is detected.

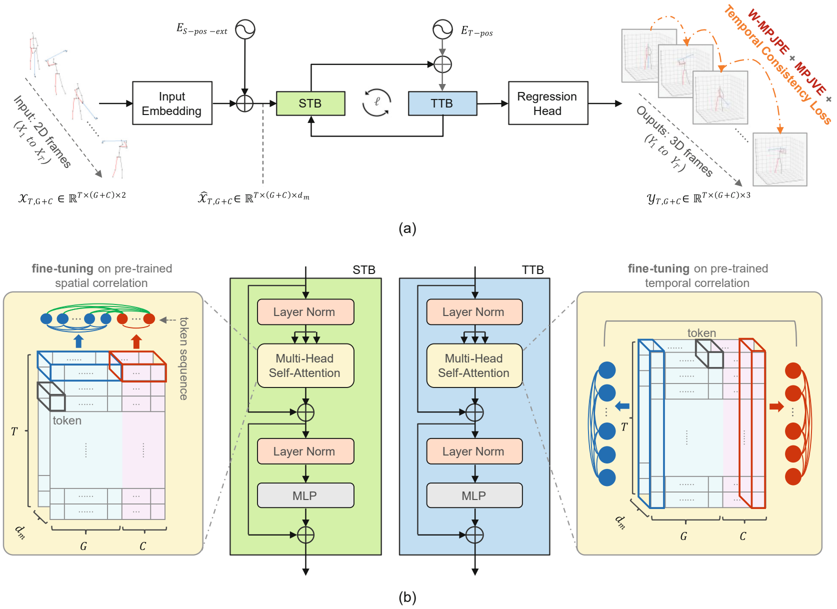
For 2D pose detection, we employ HRNet [31] and ViTPose-H [37], which are pretrained on the COCO 2017 dataset for 17 human keypoints. We also include DEKR, which is a bottom-up method, for comparison purpose. Refering to Fig. 1, we then fine-tune them into three models by *GolfSwing* (2D).

- For golfer pose, we modify the configuration file according to the Human3.6M keypoint format. The backbones of the above three models are initialized by pre-trained weights, and we retrain their keypoint heads from scratch, leading to 17 keypoints as output.
- For club pose, we also use the backbones of the above three models and load their pretrained weights. Then we modify their prediction heads for 5 club keypoints and fine-tune them by *GolfSwing* (2D). The club keypoints are shaft, hosel, heel, toe down, and toe-up.
- For golfer-with-club pose, the same fine-tuning is executed except that there are 17+5 keypoints from prediction heads as output.

For 2D-3D lifting, there has been extensive research [21,28,39,43]. We choose to fine-tune the state-of-the-art MixSTE [39]. We follow its design and extend the dimensions of the model to include 5 extra club keypoints. The process is shown in Fig. 3(a). The input to the model is a sequence of T 2D poses $\mathcal{X}_{T,G+C} \in \mathbb{R}^{T\times(G+C)\times 2}$, where G and C are the numbers of golfer and club keypoints, respectively. First, we project each keypoint to d_m dimensions, leading to a higher-dimension feature map $\hat{\mathcal{X}}_{T,G+C} \in \mathbb{R}^{T\times(G+C)\times d_m}$. Then, to preserve positional information, the extended spatial embedding matrix $E_{s\text{-}pos\text{-}ext} \in \mathbb{R}^{(G+C)\times d_m}$ and pre-trained temporal embedding matrix $E_{t\text{-}pos} \in \mathbb{R}^{T\times d_m}$ are applied. (Note that the pre-trained embedding $E_{s\text{-}pos} \in \mathbb{R}^{G\times d_m}$, which is trained on human keypoints only, can not be directly fine-tuned.) Therefore, we randomly initialize $E_{s\text{-}pos\text{-}ext}$ for retaining the positional information of both golfer and club during fine-tuning. Subsequently, $\hat{\mathcal{X}}_{T,G+C}$ will be iternately learned for l iterations between Spatial Transformer Block (STB) and Temporal Transformer Block (TTB). Finally, the dimension d_m is reduced to 3 by the Regression Head, leading to a keypoint sequence $\mathcal{Y}_{T,G+C} \in \mathbb{R}^{T\times(G+C)\times 3}$.

The modified STB and TTB transformer blocks are shown in Fig. 3(b). We follow the transformer encoders designed in [6,39,43]. STB is to learn the spatial relationships among keypoints in each frame. Frames are sent one-by-one to STB. With the pre-trained weights of MixSTE, the *multi-head self-attention* of STB already effectively preserved the spatial relation of keypoints for regular human activities. During fine-tuning, for each frame at time t, its (dimension-incremented) keypoints, denoted by $i_{t,n} \in \mathbb{R}^{d_m}, n = 1...(G+C)$, are regarded as a sequence of tokens by STB to enhance their spatial relation-capturing capability, such as inter-golfer, inter-club, and golf-club keypoints' relationships. On the other hand, the trajectories of all keypoints along the temporal dimension are also sent one-by-one to TTB. For each trajectory n, $n = 1..G+C$, its (dimension-incremented) keypoints, denoted by $i_{n,t} \in \mathbb{R}^{d_m}, t = 1...T$, are regarded as a sequence of tokens by TTB to enhance their temporal relation-capturing capability. Overall, these two blocks alternately strengthen the correlations of keypoints in spatial and temporal dimensions, respectively.

This model is fine-tuned end-to-end in a supervised manner. We adopt the same loss functions: Weight Mean Per Joint Position Error (W-MPJPE) L_w and Mean Per Joint Velocity Error (MPJVE) L_v [28]. Additionally, we adopt Temporal Consistency Loss (TCLoss) L_c to improve motion smoothness [9].


Input: 2D frames
(X_1 to X_T)
$\mathcal{X}_{T,G+C} \in \mathbb{R}^{T\times(G+C)\times 2}$
Input Embedding
$E_{S-pos-ext}$
$\hat{\mathcal{X}}_{T,G+C} \in \mathbb{R}^{T\times(G+C)\times d_m}$
STB
ℓ
TTB
E_{T-pos}
Regression Head
Ouputs: 3D frames
(Y_1 to Y_T)
$\mathcal{Y}_{T,G+C} \in \mathbb{R}^{T\times(G+C)\times 3}$
W-MPJPE + MPJVE + Temporal Consistency Loss
fine-tuning on pre-trained spatial correlation
token sequence
token
T
d_m
G
C
Layer Norm
Multi-Head Self-Attention
Layer Norm
MLP
fine-tuning on pre-trained temporal correlation

(a)

(b)

Fig. 3. (a) Extension of MixSTE for 2D-3D lifting and (b) extension of STB and TTB transformer blocks to include club keypoints.

3.3 GolfPose Inference Model

GolfPose is built upon the above fine-tuned models. As shown in Fig. 1, it accepts a RGB frame sequence of length (T, H, W, C) as input. If one chooses to process golfer and club separately, we need to identify from each frame a golfer bounding box $G_b = (p_x, p_y, p_w, p_h, score)$ and a club bounding box $C_b = (c_x, c_y, c_w, c_h, score)$. Then, P_b and C_b are passed to the golfer and the club pose models, respectively. Then golfer and club keypoints of all T frames are stacked into tensors of $(T, 17, 2)$ and $(T, 5, 2)$, respectively, which are then concatenated into a $(T, 22, 2)$ tensor. If one chooses to process golfer and club jointly, the process is similar, except that there is only one bounding box per frame and we directly derive a $(T, 22, 2)$ tensor. In either case, the concatenated tensor is fed into the 2D-3D lifter. With joint golfer-with-club information, we shall show that the lifter can better leverage the spatial-temporal correlations of keypoints and thus achieve much higher accuracy.

4 Performance Evaluation

4.1 Implementation Details

As mentioned earlier, our 2D/3D models are pre-trained on the COCO 2017 dataset and the Human3.6M dataset, respectively, and then fine-tuned on *GolfSwing* 2D/3D. For *GolfSwing*, the training set (S1-S4) consists of 13,782 images,

Table 1. Comparisons of 3D pose estimation models on our *GolfSwing* 3D dataset.

Strategy	w/o Fine-tuned	with Fine-tuned	
	Golfer	Golfer	Club
VideoPose3D [28] (N=17, T=243)	134.8	52.0	-
Attention3D [21] (N=17, T=243)	149.7	46.8	-
PoseFormer [43] (N=17, T=81)	**107.8**	40.3	-
MixSTE [39] (N=17, T=243)	109.4	**35.6**	-
GolfPose-3D(GC) (N=22, T=243)	-	**32.3**	62.8

2D/3D keypoints, and 27,564 bounding box annotations, while the test set (S5, S6) consists of 3,956 images, 2D/3D keypoints, and 7,912 bounding box annotations. For object detectors, the evaluation metric is $mAP@IoU$. During fine-tuning, we set a batch size of 8 and train the models for 30 epochs. The optimizer is SGD, and the learning rate is set to 2.5e-3. The computing environments are: CPU i7-12700K, GPU GeForce RTX 3090*2, CUDA 11.6, PyTorch 1.12.1, and mmdetection [4] version 3.1.0.

For 2D pose models, the evaluation metric is $mAP@OKS$. The computing environments are the same but with additional mmpose [5] version 1.3.0. During fine-tuning, we set the batch size to 16 and train the models for 20 epochs. We use Adam optimizer, with a learning rate of 1e-4. For the golfer's 17 keypoints, we assign different weights [1.0, 1.0, 1.2, 1.5, 1.0, 1.2, 1.5, 1.0, 1.0, 1.0, 1.0, 1.0, 1.2, 1.5, 1.0, 1.2, 1.5] to them when calculating MSE loss. For the club's 5 keypoints, we assign weights [1.6, 1.9, 2.0, 2.0, 2.0] to them. The golfer-with-club's keypoionts are given weights similarly.

To fine-tune MixSTE, in addition to extending to 22 keypoints, we perform data augmentation on *GolfSwing* to enhance robustness. We rotate each 3D pose by 90 degrees and project it onto the 2 RGB cameras. So the dataset quadrupled, effectively rendering additional perspectives of 2D poses. We divide keypoints into five groups (head, torso, upper limbs, lower limbs, and club) and define the weight vector $W = [1.5, 1, 2.5, 4, 4]$. The frame length $T = 243$ and the Adam optimizer is employed with a learning rate of 4.0e-5 and a decay of 0.98 per epoch. The batch size is 512 and the model is fine-tuned for 60 epochs. The computing environments are: CPU i7-12700K, GPU GeForce RTX 3090*2, CUDA 11.6, and PyTorch 1.10.1.

4.2 Performance Comparison

Quantitative Results. We compare *GolfPose* against four 3D models VideoPose3D [28], Attention3D [21], PoseFormer [43], and MixSTE [39] on the *GolfSwing* dataset. Among them, MixSTE is the current state-of-the-art in Human3.6M. We use the 2D pose ground truth as input to compare the predicted 3D poses by the MPJPE metric. We use the default hyper-parameters of

Table 2. Comparisons of 2D pose models on our *GolfSwing* 2D dataset. (G, C, and GC mean fine-tuning for golfer only, for club only, and for both, respectively. "Metric" means the range of keypoints in calculating AP and AR.)

Model	Source model	Metric	AP	AP^{50}	AP^{75}	AR	AR^{50}	AR^{75}
GolfPose-2D(G)	HRNet	AP_{golfer} AR_{golfer}	0.884	**1.000**	**1.000**	0.887	**1.000**	**1.000**
GolfPose-2D(GC)			0.899	**1.000**	**1.000**	0.916	**1.000**	**1.000**
GolfPose-2D(G)	ViTPose-H		0.887	**1.000**	**1.000**	0.898	**1.000**	**1.000**
GolfPose-2D(GC)			0.901	**1.000**	**1.000**	0.915	**1.000**	**1.000**
GolfPose-2D(G)	DEKR		0.869	**1.000**	0.898	0.888	**1.000**	0.904
GolfPose-2D(GC)			**0.917**	**1.000**	0.968	**0.927**	**1.000**	0.973
GolfPose-2D(C)	HRNet	AP_{club} AR_{club}	0.857	0.990	0.947	0.882	0.997	0.957
GolfPose-2D(GC)			0.949	**1.000**	0.990	0.955	**1.000**	0.999
GolfPose-2D(C)	ViTPose-H		0.870	0.990	0.948	0.887	0.996	0.953
GolfPose-2D(GC)			0.942	**1.000**	0.990	0.956	**1.000**	0.998
GolfPose-2D(C)	DEKR		0.858	0.990	0.946	0.888	0.999	0.951
GolfPose-2D(GC)			**0.977**	**1.000**	**1.000**	**0.982**	**1.000**	**1.000**
GolfPose-2D(GC)	HRNet	$AP_{golfer-club}$ $AR_{golfer-club}$	0.915	**1.000**	**1.000**	0.930	**1.000**	**1.000**
	ViTPose-H		0.925	**1.000**	**1.000**	0.930	**1.000**	**1.000**
	DEKR		**0.942**	**1.000**	**1.000**	**0.945**	**1.000**	**1.000**
HRNet [31]	-	AP_{limb} AR_{limb}	0.701	**1.000**	0.948	0.731	**1.000**	0.954
GolfPose-2D(G)	HRNet		0.918	**1.000**	**1.000**	0.939	**1.000**	**1.000**
GolfPose-2D(GC)	HRNet		**0.956**	**1.000**	**1.000**	**0.962**	**1.000**	**1.000**
ViTPose-H [37]	-		0.706	**1.000**	**1.000**	0.730	**1.000**	**1.000**
GolfPose-2D(G)	ViTPose-H		0.936	**1.000**	**1.000**	0.947	**1.000**	**1.000**
GolfPose-2D(GC)	ViTPose-H		0.941	**1.000**	**1.000**	0.948	**1.000**	**1.000**
DEKR [8]	-		0.669	**1.000**	0.979	0.689	**1.000**	0.988
GolfPose-2D(G)	DEKR		0.877	**1.000**	0.868	0.887	**1.000**	0.871
GolfPose-2D(GC)	DEKR		0.918	**1.000**	0.927	0.924	**1.000**	0.935

these four models during fine-tuning. As Table 1 shows, these four models all improve significantly after fine-tuning, implying the contribution of *GolfSwing*. After fine-tuning, MixSTE performs the best. Encompassing club information, *GolfPose* generates $N = 22$ keypoints and outperforms the other methods, which only yield $N = 17$ golfer keypoints. This indicates that including object is helpful for pose estimation. After fine-tuning, *GolfPose* achieves the lowest MPJPE of 32.3 mm for golfer keypoints. In fact, the club's MPJPE=62.8 mm because its fast-moving nature causes blurry effects. Even under such a condition, it still proves the importance of including club for golfer pose estimation.

Next, we consider the 2D pose estimation results, including golfer-only, club-only, and golfer-with-club cases. Table 2 presents two types of results: HRNet and ViTPose-H represent the top-down approach, and DEKR represents the bottom-up approach. If we fine-tune for golfer only (G) or for club only (C) by *GolfSwing* 2D, ViTPose-H performs the best with mAP=0.887 and 0.870, respectively (underlined). If we fine-tune for both golfer and club (GC), all models are further improved after fine-tuning. DEKR achieves the highest mAP of 0.917 in golfer's keypoints, of 0.977 in club's keypoints, and of 0.942 in all keypoints (boldface). These results indicate that including club benefits golfer keypoint detection, and reversely including golfer benefits club keypoint detection.

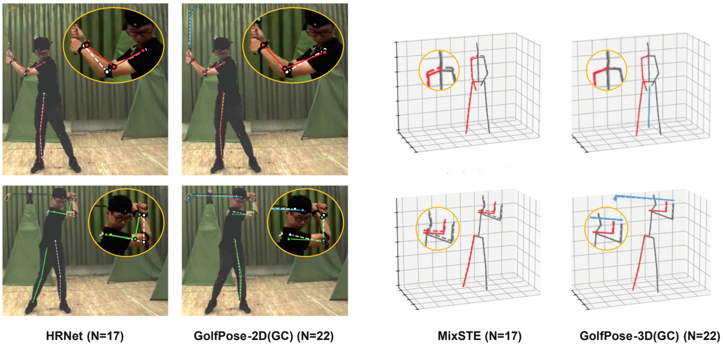


Fig. 4. Qualitative comparisons on subject S5. (GT=dashed line; prediction=solid line; red=right hand; green=left hand)

Finally, we compare the pre-trained and the fine-tuned models. Since COCO and Human3.6M define skeleton differently, we have to take the 12 common keypoints between them (which include shoulders, elbows, wrists, hips, knees, and ankles). The results are shown in the last section of Table 2. For each source model (HRNet, ViTPose-H, and DEKR), our fine-tuned *GolfPose* does improve mAP significantly. Overall, using golfer-with-club data to fine-tune HRNet performs the best, achieving mAP= 0.956. This implies the value of *GolfSwing* that makes 2D pose estimation more stable and accurate, which can further contribute to the subsequent 2D-3D lifter.

Qualitative Results. Fig. 4 presents some qualitative results. The visualization is from S5 of *GolfSwing*. The results show the improvement from the pre-trained model to the fine-tuned model. Notably, even when hands are partially occluded, *GolfPose*-2D and -3D can still detect keypoints quite accurately.

Inference speed. Regarding inference speed, *GolfPose* mainly involves a fine-tuned 2D golfer pose model, a fine-tuned 2D club pose model, and a fine-tuned 2D-3D lifter. The first two models, when running on an i5-12500 CPU and GeForce GTX 1080 Ti GPU, achieve 27.25 and 27.3 FPS, respectively. The third model, when running on an i5-13400 CPU with a GeForce RTX 3060 GPU, reaches 6.67 FPS.

Object Detection. Table 3 compares the case of detecting golfer and club separately and the case of detecting them jointly. We test two object detectors: Faster R-CNN and YOLOX-s. There is clear advantage of detecting them jointly. Contrary to intuition, when each individual object's detection is low, jointly detecting them helps improve detection rate. With joint detection, YOLOX-s outperforms Faster R-CNN. Additionally, YOLOX-s boasts a higher inference speed of 88.84 FPS compared to Faster R-CNN's 14.19 FPS. Therefore, we adopt YOLOX-s as our object detector.

4.3 Ablation Study

Number of club keypoints. In Table 4, we further consider the effect of the number of club keypoints. We denote the method by $17+i$, where $i = 0..5$ represents the number of club keypoints (when $0 < i < 5$, we choose keypoints c1 to ci in Fig. 2). When we start to add club keypoint, we observe significant improvement on both golfer and golfer-with-club detection accuracy (from error=35.6 mm to 29-32 mm for golfer). However, adding more keypoints results in slight increases of error. We suspect the reason to be the relative slower movement of the grip part as opposed to the much faster movement of the head part of club. As mentioned earlier, it is more difficult to detect fast-moving keypoint. Therefore, when the value of i increases, these keypoints (of relative lower accuracy) also confuse our model.

Table 3. Ablation study on separate and joint object detection.

Model	Datasets	Class	AP	AP^{50}	AP^{75}
Faster R-CNN (ResNet50-FPN)	Coco + *GolfSwing*	Golfer	0.940	1.000	1.000
		Club	0.896	1.000	0.989
		G-w-C	0.970	1.000	0.990
YOLOX-s (CSPDarknet)	Coco + *GolfSwing*	Golfer	0.920	1.000	1.000
		Club	0.911	1.000	0.998
		G-w-C	0.984	1.000	1.000

Table 4. Ablation study on the number of club keypoints.

MPJPE (mm)	Number of keypoints					
	17+0	17+1	17+2	17+3	17+4	17+5
Golfer	35.6	29.5	30.5	30.8	32.3	32.3
Club	-	50.9	59.6	62.9	61.4	62.8
Overall	35.6	30.7	33.6	35.6	37.9	39.2

Table 5. Ablation study on fine-tuning from Human3.6M.

MPJPE (mm)	Train from scratch	Fine-tuning
Golfer	48.5	32.3
Club	112.9	62.8
Overall	63.2	39.2

Effect of fine-tuning. We consider the same structure of *GolfPose* that is trained from scratch on *GolfSwing* for 80 epochs (i.e., without using the pre-trained weights from MixSTE). From Table 5, it validates the benefit of the pretrained weights from MixSTE (which reduces error from 48.5 mm to 32.3 mm for golfer). That is, a large amount of information is carried over from the pre-trained weights obtained from Human3.6M. Even for club, the error is reduced from 112.9 mm to 62.8 mm.

5 Conclusions

This work contributes in deriving the *GolfSwing* dataset, which includes keypoint ground truth of 2D and 3D golf swing actions. It also contributes in deriving the *GolfPose* framework, which can be fine-tuned from existing object detection and pose estimation models, for inferring golfer-with-club keypoints simultaneously. The results imply that including auxiliary objects, such as club, with even very few keypoints of a small dataset can improve human pose estimation significantly. Nonetheless, detecting club poses in complex scenes is a challenge. Future improvement on club pose estimation may further improve overall performance. This approach can be extended to other sports, such as baseball, cricket, badminton, and tennis, where players have an object at hand.

References

1. Baumgartner, T., Klatt, S.: Monocular 3d human pose estimation for sports broadcasts using partial sports field registration. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5108–5117 (2023)
2. Cao, Z., Simon, T., Wei, S.E., Sheikh, Y.: Realtime multi-person 2d pose estimation using part affinity fields. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 7291–7299 (2017)
3. Captury: Captury motion systems. https://captury.com/ (2013), accessed: 2023-06-19
4. Chen, K., Wang, J., Pang, J., Cao, Y., Xiong, Y., Li, X., Sun, S., Feng, W., Liu, Z., Xu, J., Zhang, Z., Cheng, D., Zhu, C., Cheng, T., Zhao, Q., Li, B., Lu, X., Zhu, R., Wu, Y., Dai, J., Wang, J., Shi, J., Ouyang, W., Loy, C.C., Lin, D.: MMDetection: Open mmlab detection toolbox and benchmark. arXiv preprint arXiv:1906.07155 (2019)
5. Contributors, M.: Openmmlab pose estimation toolbox and benchmark. https://github.com/open-mmlab/mmpose (2020)
6. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. arXiv preprint arXiv:2010.11929 (2020)
7. Ge, Z., Liu, S., Wang, F., Li, Z., Sun, J.: Yolox: Exceeding yolo series in 2021. arXiv preprint arXiv:2107.08430 (2021)
8. Geng, Z., Sun, K., Xiao, B., Zhang, Z., Wang, J.: Bottom-up Human Pose Estimation via Disentangled Keypoint Regression. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 14676–14686 (2021)

9. Hossain, M.R.I., Little, J.J.: Exploiting Temporal Information for 3D Human Pose Estimation. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 68–84 (2018)
10. Ingwersen, C.K., Mikkelstrup, C., Jensen, J.N., Hannemose, M.R., Dahl, A.B.: SportsPose: A Dynamic 3D Sports Pose Dataset. In: Proceedings of the IEEE/CVF International Workshop on Computer Vision in Sports (2023)
11. Ionescu, C., Papava, D., Olaru, V., Sminchisescu, C.: Human3.6m: Large Scale Datasets and Predictive Methods for 3D Human Sensing in Natural Environments. IEEE Transactions on Pattern Analysis and Machine Intelligence **36**(7), 1325–1339 (2013)
12. Kanazawa, A., Black, M.J., Jacobs, D.W., Malik, J.: End-to-end Recovery of Human Shape and Pose. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 7122–7131 (2018)
13. Kim, T.T., Zohdy, M.A., Barker, M.P.: Applying Pose Estimation to Predict Amateur Golf Swing Performance using Edge Processing. IEEE Access **8**, 143769–143776 (2020)
14. Lee, K.J., Ryou, O., Kang, J.: Quantitative Golf Swing Analysis based on Kinematic Mining Approach. Korean Journal of Sport Biomechanics **31**(2), 87–94 (2021)
15. Liao, C.C., Hwang, D.H., Koike, H.: AI Golf: Golf Swing Analysis Tool for Self-Training. IEEE Access **10**, 106286–106295 (2022)
16. Liao, C.C., Hwang, D.H., Wu, E., Koike, H.: AI Coach: A Motor Skill Training System using Motion Discrepancy Detection. In: Proceedings of the Augmented Humans International Conference. pp. 179–189 (2023)
17. Lin, K., Wang, L., Liu, Z.: End-to-end Human Pose and Mesh Reconstruction with Transformers. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 1954–1963 (2021)
18. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft COCO: Common Objects in Context. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 740–755. Springer International Publishing (2014)
19. Liu, J., Saquib, N., Chen, Z., Kazi, R.H., Wei, L.Y., Fu, H., Tai, C.L.: PoseCoach: A Customizable Analysis and Visualization System for Video-based Running Coaching. In: IEEE Transactions on Visualization and Computer Graphics. pp. 1–14 (2022)
20. Liu, P., Wang, J.H.: MonoTrack: Shuttle Trajectory Reconstruction From Monocular Badminton Video. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). pp. 3513–3522 (2022)
21. Liu, R., Shen, J., Wang, H., Chen, C., Cheung, S.c., Asari, V.: Attention Mechanism Exploits Temporal Contexts: Real-time 3D Human Pose Reconstruction. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 5064–5073 (2020)
22. Timo von Marcard, T., Henschel, R., Black, M.J., Rosenhahn, B., Pons-Moll, G.: Recovering Accurate 3D Human Pose in The Wild Using IMUs and a Moving Camera. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 601–617 (2018)
23. McNally, W., Vats, K., Pinto, T., Dulhanty, C., McPhee, J., Wong, A.: Golfdb: A Video Database for Golf Swing Sequencing. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPR). pp. 0–0 (2019)

24. Mehta, D., Rhodin, H., Casas, D., Fua, P., Sotnychenko, O., Xu, W., Theobalt, C.: Monocular 3D Human Pose Estimation In The Wild Using Improved CNN Supervision. In: Proceedings of the International Conference on 3D Vision (3DV) (2017)
25. Mohamed, A., Chen, H., Wang, Z., Claudel, C.: Skeleton-graph: Long-term 3D Motion Prediction from 2D Observations using Deep Spatio-temporal Graph CNNs. arXiv preprint arXiv:2109.10257 (2021)
26. Mun, F., Suh, S.W., Park, H.J., Choi, A.: Kinematic Relationship Between Rotation of Lumbar Spine and Hip Joints during Golf Swing in Professional Golfers. Biomed. Eng. Online **14**, 1–10 (2015)
27. Nonaka, N., Fujihira, R., Nishio, M., Murakami, H., Tajima, T., Yamada, M., Maeda, A., Seita, J.: End-to-End High-Risk Tackle Detection System for Rugby. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). pp. 3550–3559 (2022)
28. Pavllo, D., Feichtenhofer, C., Grangier, D., Auli, M.: 3D Human Pose Estimation in Video with Temporal Convolutions and Semi-supervised Training. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 7753–7762 (2019)
29. Ren, S., He, K., Girshick, R., Sun, J.: Faster r-cnn: Towards real-time object detection with region proposal networks. IEEE Transactions on Pattern Analysis and Machine Intelligence (Jun 2017)
30. Roetenberg, D., Luinge, H., Slycke, P., et al.: Xsens MVN: Full 6DOF Human Motion Tracking using Miniature Inertial Sensors. Xsens Motion Technologies BV, Tech. Rep **1**, 1–7 (2009)
31. Sun, K., Xiao, B., Liu, D., Wang, J.: Deep High-resolution Representation Learning for Human Pose Estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 5693–5703 (2019)
32. Tekin, B., Rozantsev, A., Lepetit, V., Fua, P.: Direct Prediction of 3D Body Poses from Motion Compensated Sequences. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 991–1000 (2016)
33. Trumble, M., Gilbert, A., Malleson, C., Hilton, A., Collomosse, J.: Total Capture: 3D Human Pose Estimation Fusing Video and Inertial Sensors. In: Proceedings of British Machine Vision Conference. pp. 1–13 (2017)
34. Vicon: Motion Capture. https://www.vicon.com/ (1984), accessed: 2023-08-07
35. Wu, Y., Kirillov, A., Massa, F., Lo, W.Y., Girshick, R.: Detectron2. https://github.com/facebookresearch/detectron2 (2019)
36. Xiao, B., Wu, H., Wei, Y.: Simple Baselines for Human Pose Estimation and Tracking. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 466–481 (2018)
37. Xu, Y., Zhang, J., Zhang, Q., Tao, D.: Vitpose: Simple vision transformer baselines for human pose estimation. In: Advances in Neural Information Processing Systems (2022)
38. Yu-Hui, C., Ard, O., Francois, B., Andrew, B., Vijay, S.: MoveNet. https://www.tensorflow.org/hub/tutorials/movenet (2021)
39. Zhang, J., Tu, Z., Yang, J., Chen, Y., Yuan, J.: MixSTE: Seq2seq Mixed Spatio-Temporal Encoder for 3D Human Pose Estimation in Video. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 13232–13242 (2022)
40. Zhang, Z.: A Flexible New Technique for Camera Calibration. IEEE Trans. Pattern Anal. Mach. Intell. **22**(11), 1330–1334 (2000)

41. Zhao, L., Peng, X., Tian, Y., Kapadia, M., Metaxas, D.N.: Semantic Graph Convolutional Networks for 3D Human Pose Regression. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 3425–3435 (2019)
42. Zheng, C., Wu, W., Chen, C., Yang, T., Zhu, S., Shen, J., Kehtarnavaz, N., Shah, M.: Deep Learning-based Human Pose Estimation: A Survey. arXiv preprint arXiv:2012.13392 (2020)
43. Zheng, C., Zhu, S., Mendieta, M., Yang, T., Chen, C., Ding, Z.: 3D Human Pose Estimation With Spatial and Temporal Transformers. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 11656–11665 (2021)
44. Zheng, N., Barrentine, S., Fleisig, G., Andrews, J.: Swing Kinematics for Male and Female Pro Golfers. Int. J. Sports Med. **29**(12), 965–970 (2008)

Fusing Multimodal Streams for Improved Group Emotion Recognition in Videos

Deepak Kumar(✉), Piyush Dhamdhere, and Balasubramanian Raman

Department of Computer Science and Engineering,
Indian Institute of Technology Roorkee, Roorkee 247667, India
{d_kumar,piyush_d,bala}@cs.iitr.ac.in

Abstract. Recognizing social cues and emotions is vital for navigating daily interactions, understanding emotions in conversations, interpreting body language in meetings, and supporting friends in difficult situations. This work focuses on analyzing group-level emotions in videos captured in natural settings, marking an attempt at multimodal group-level emotion analysis. Automatic group emotion recognition is pivotal for understanding complex human-human interactions. Group emotion recognition in videos presents several challenges because existing work predominantly focuses either on individual emotion recognition or group emotion analysis in static images. To address this challenge, we introduce a deep-learning-based multimodal fusion model that integrates diverse modalities, including audio, video, and scene. Feature extraction employs advanced models like TimeSformer for video description and wav2vec2.0 for audio analysis. All the experiments are conducted on the VGAF dataset. Our key findings include: (1) Multimodal approaches outperform their unimodal counterparts, (2) Experimental results confirm the superior performance of proposed approach compared to benchmark methods on the given dataset, and (3) There is a strong correlation between modalities and respective emotions.

Keywords: Affective computing · Group level affect recognition · Human behavior analysis

1 Introduction

Understanding emotions is crucial in interactions between humans and computers. The field of emotion recognition research is expanding quickly because of the potential applications for improved human-computer interfaces and automated services that react instantly to the user's or client's emotions [13]. Group emotion prediction is a very first step towards the development of artificial intelligence (AI) systems that will be able to understand complicated human connections and facilitate high-level human interaction [40]. Visual cues are insufficient to

Supported by the Ministry of Education (MoE) INDIA with reference grant number: OH-3123200428

A. Antonacopoulos et al. (Eds.): ICPR 2024, LNCS 15321, pp. 403–418, 2025.
https://doi.org/10.1007/978-3-031-78305-0_26

capture the complexity of human interactions, as collective emotion is a reflection of intentions, behaviours, and relationships within the group. Recognizing emotions in groups using both audio and video is tough because it's hard to collect all the necessary information to understand emotions properly.

Humans are able to understand a video's context in addition to the audiovisual content that is shown, based on the mechanism that converts the conceptual information that has been gathered in the brain into a visual cognition process [8,23,24]. Thus, we believe that acquiring valuable insights into understanding collective emotions can be facilitated by exploring various aspects of human relationships. Unlike machines, humans can grasp emotions more easily because we connect what we see and hear with our memories, making it easier to understand the context. Recent developments in emotional technology aim to make our devices capable of understanding emotions as humans do. Recent interest in group-level emotion recognition in real-world settings has surged due to innovative data collection methods and multimodal capabilities. While the social psychology community has long been intrigued by group-level emotion recognition [25], the importance lies in understanding social identity and individuals' interactions within their social context. Beyond social psychology, various disciplines such as Education and healthcare are also fascinated by group emotions.

Automatic group-level affect analysis presents greater challenges due to various factors. These include the potential occlusion of individuals in group scenarios due to camera placement, variations in label assignments by different human annotators focusing on different group segments, and the influence of multiple individuals on each other's emotional states during interactions [20]. To work with these challenges, we employed multiple cues, including speech, pose, frames, and overall video representation, for group-level affect analysis. Overall, we make the following contributions:

- We introduce a novel multimodal decision fusion model, leveraging four streams of data, each corresponding to a distinct feature type: audio, pose, frame, and video representation, to classify group emotions in a multiclass setting. Additionally, we investigate various methods for integrating information from these video features.
- Our experiments reveal that multimodal fusion involving all four features: audio, pose, frame, and video is highly predictive of emotion classes. We also show that fusion approaches outperform their unimodal ones preserving the cue complementarity of feature space.
- We have tested proposed approach on the public dataset: VGAF which classify emotions into one of three categories. Experimental findings confirms a performance improvement over benchmark approaches.
- We have also investigated the correlation between emotion classes and the corresponding feature space.

The subsequent sections of the paper are organized as follows: Section 2 delves into previous research works in this domain. Following that, Section 3 discuss the architecture and methodology employed. Section 4 covers the experimental outcomes. At the end, the conclusion and future work is discussed.

2 Background and Related Work

Recognizing emotions from images, particularly through facial cues [21], or via video analysis [1], has long been an established field of research. However, the exploration of group-level affect recognition is a rapidly evolving research area. Recent advancements in machine learning technology have facilitated the ability to recognize emotions in real-world scenarios, commonly referred to as "in-the-wild" settings. Notably, current studies have been focusing on the challenging task of predicting emotions at the group level in video contexts [36,37]. Initially, researchers prioritized unimodal approaches [28,30], utilizing single cues only, but now the majority employ multimodal inputs [5,11]. Various models have been proposed to achieve group affect recognition, ranging from classic machine learning-based approaches [17] to deep learning-based methods and more recent attention-based techniques [42].

Several notable works contribute to the field: the work by Zhang et al. [41], which introduces ERLDK, a RL (Reinforcement Learning) based model for multimodal emotion recognition in conversation videos that utilize historical utterances, a DDQN with GRU layers, and domain knowledge to refine recognition across diverse datasets. Liu et al. [22] introduce a hybrid network by integrating audio, facial emotion, environmental object statistics (EOS), and video streams. The EOS method effectively leverages facial expressions, environmental context, audio, and temporal features for Audio-Visual group emotion recognition. Also, Augusma et al. [2] propose a multimodal model with video and audio branches employing cross-attention by utilizing a fine-tuned Vision Transformer (ViT) architecture for video and convolutional neural network (CNN) blocks feeding Mel-spectrograms into a transformer encoder for audio. Another work [38] presents the K-injection audiovisual network, capturing both explicit and implicit knowledge in emotion recognition. Additionally, in work by Guo et al. [14], recent deep neural models trained on various cues including facial expressions, scene context and skeletal features are integrated for group emotion classification. Also, Zhao et al. [42] propose the VisualAudio Attention Network (VAANet), an end-to-end approach for video emotion recognition that integrates different attention mechanisms such as spatial and channel-wise attention into an audio-visual 2D/3D CNN network.

On the other hand, along with deep learning models, Pinto et al. [29] also utilize machine learning models for group affect classification by integrating a pretrained Inflated ResNet-50 model for visual cues, processes audio features using a Bidirectional Long Short-Term Memory (Bi-LSTM) network, and employs a support vector machine classifier for classification. While, Dhall et al. [17] utilize a global alignment kernel to explicitly measure the distance between two images and introduce SVM-CGAK, a support vector machine approach for group- emotion recognition. Balaji et al. [4] also employ SVM-based classification approach by incorporating low-level and mid-level components, alongside deep neural models for feature extraction to classify group emotion. Also, in one of the interesting work, Pan et al. [27] proposed a Random Forest based approach adopted for fusing the fused features from various CNNs.

3 Methodology

3.1 Problem Formulation

Let $\mathbf{X}$ denote the input video, where $\mathbf{X} \in \mathbb{R}^{F \times H \times W \times C}$. Here, F represents the number of frames each of size: height $(H) \times$ width (W), and C encompasses the number of channels in the video frame (see Figure 3). The task involves analyzing a 5-second video $(\mathbf{X})$ captured in natural settings featuring a group of N individuals (where N>1). The goal is to categorize the overall emotion of the entire video into one of three classes: Positive, Negative, or Neutral using multiclass classification. Furthermore, the objective extends to investigating the correlation between the utilized features and the available emotion class.

3.2 Features Extraction

We have extracted different features from the videos, such as audio, frame, pose, and video features.

1. **Audio Features:** Audio features, including spectrograms, MFCCs, and pitch [9,31], are numerical representations extracted from audio signals. We have utilized wave2vec2.0 [3] for audio feature extraction due to its ability to learn dense and semantically rich representations from sequential data. This resulted in a 1024-dimensional audio embedding vector for each video.
2. **Pose Features:** Pose estimation requires detecting key points representing body parts' positions and orientations in images. YOLOv8 [19] introduces specialized models for this, adept at accurately identifying key points across diverse contexts. Inspired by this, we have utilized YOLOv8 on our video frames, Let Yolo give ρ number of points as: $\{(x_1, y_1), (x_2, y_2), \ldots, (x_\rho, y_\rho)\}$ of a person. In a frame, there are 'm' individuals (where 'm' may vary from frame to frame). The average of each detected point (x, y) for each frame is then calculated as:

$$X = (x_1 + x_2 ....x_m)/m \tag{1}$$

$$Y = (y_1 + y_2 ....y_m)/m \tag{2}$$

The mean features extracted from each frame are stacked to form the final feature representation. Specifically, YOLOv8 detects a total of 17 key points, but for our specific task, only 8 points - including wrist and hip points, which are consistently visible in each video frame - is selected. This results in a 16-dimensional vector. Furthermore, the count of people present in the video is appended, yielding a 17-dimensional feature set. To ensure uniform temporal sampling and comprehensive coverage, we selected 12 representative frames from each 5-second video, with frame rates ranging from 13 to 30 fps (resulting in at least 65 frames), using 12 equally spaced frame indices. This choice balances computational efficiency and pose variation. By sampling at regular intervals, we capture different stages of movement, effectively analyzing human poses and providing a broad view of posture changes. Considering 12 representative frames per video, results in total 204-dimensional feature vector for each video.

3. **Frame Features:** A video consists of a sequence of frames, each capturing the entire scene at a particular moment. To capture temporal information, we have concentrated on the video's temporal dimension by extracting frame-level features. We utilize the OpenCV [16] library in Python to extract frames from the input video. Additionally, after taking inspiration from [18], a compact spatio-temporal network to extract frame-level features is used. The spatio-temporal network, shown in Figure 1, incorporates several layers, including LSTM-Conv2D and average pooling, which results in 30-dimensional features for each video. Here, the time-distributed layer is utilized to apply the pre-trained ResNet-50 model [15], which was trained on ImageNet, directly on each extracted frame using a time-distributed approach. The output from this backbone was then passed through three 2D convolutional LSTM layers, each comprising 10 filters with a kernel size of 3×3, followed by an average pooling layer. Subsequently, the outputs from these three layers were concatenated.
4. **Video Features:** Several video understanding models [12,35,39] are available, but TimeSformer [6] architecture exclusively utilizes self-attention across spatial and temporal dimensions, bypassing convolutional methods. This approach enables direct spatiotemporal feature learning from sequences of frame-level patches. Motivated by these properties, we employ TimeSformer for comprehensive video feature extraction. The TimeSformer takes video clip $\mathbf{X} \in \mathbb{R}^{F \times H \times W \times C}$ as input and result in a 768-dimensional feature vector for each video.

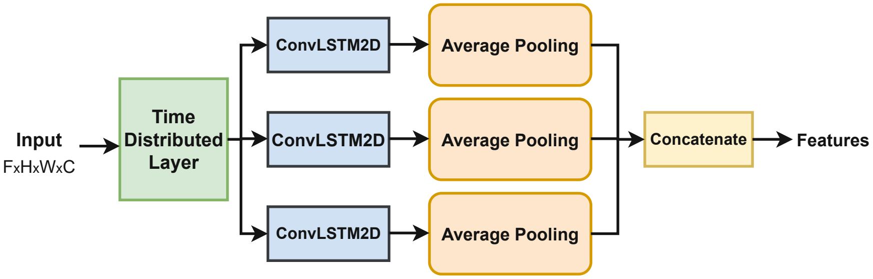


Fig. 1. Architecture of the network used in frame feature extraction.

3.3 Network architectures

Figure 2 illustrates the architecture of our proposed system, which includes the feature extraction module, multichannel module, and decision fusion module. In

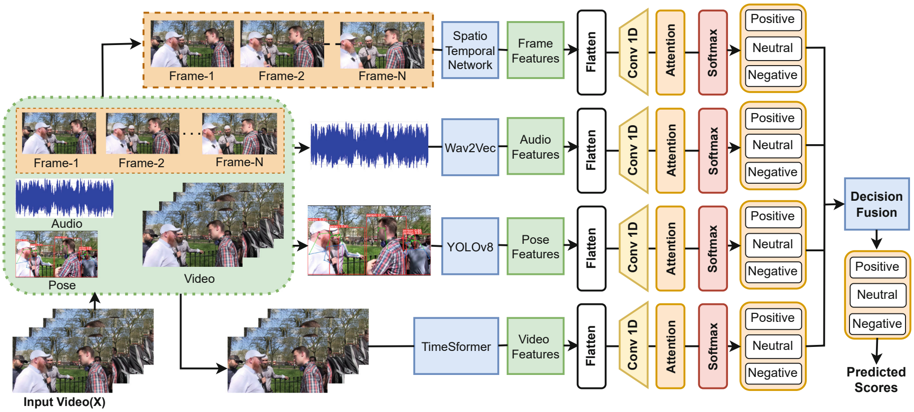


Fig. 2. Overall architecture of the proposed methodology. The network contains four deep-learning models to extract the audio, frame, pose, and frame features.

the domain of video data, various fusion strategies [33,34,39], exist, with our emphasis specifically on decision fusion, chosen based on the ablation studies detailed in section 4.4. Our approach incorporates three architectures: 1) Unimodal, 2) Bimodal/Trimodal, and 3) Multimodal decision fusion.

Multimodal decision fusion: Let F_{Audio}, F_{Frame}, F_{Pose}, and F_{Video} are features coming from Audio, Frame, Pose, and Video modality, respectively.

$$\begin{aligned} \mathbf{X}_{Audio} &= \text{vec}(F_{Audio}), \quad \mathbf{X}_{Frame} = \text{vec}(F_{Frame}), \\ \mathbf{X}_{Pose} &= \text{vec}(F_{Pose}), \quad\ \ \mathbf{X}_{Video} = \text{vec}(F_{Video}), \end{aligned} \tag{3}$$

Here, vec(.) is the function that converts the input into the dimensions required for the 1D-CNN. Let F_{Audio} is a vector given to vec(.), below operations will be performed by the vec(.) function:

$$\mathbf{X}_{Flatten} = F_{Audio}.Flatten(), \tag{4}$$

$$\mathbf{X}_{Audio} = X_{Flatten}.reshape(1, -1), \tag{5}$$

Here, the reshape function adds one extra dimension to the given input vector. Similar formulations apply to Frame, Pose, and Video features to get X_{Frame}, X_{Pose}, and X_{Video}.

For each feature modality, 1D-CNN (Convolutional Neural Networks) is applied to capture local patterns. Let $\mathcal{C}(\cdot; \theta, \delta_f, \phi)$ denote a convolutional operation with θ as the filter weights, δ_f as the filter size, and ϕ as the activation function. For each feature type the convolutional layers are defined as:

$$\mathbf{E}_{Audio} = \mathcal{C}(\mathbf{X}_{Audio}; \theta_a, \delta_f, \text{ReLU}), \tag{6}$$

$$\mathbf{E}_{Frame} = \mathcal{C}(\mathbf{X}_{Frame}; \theta_f, \delta_f, \text{ReLU}), \tag{7}$$

$$\mathbf{E}_{Pose} = \mathcal{C}(\mathbf{X}_{Pose}; \theta_p, \delta_f, \text{ReLU}), \tag{8}$$

$$\mathbf{E}_{Video} = \mathcal{C}(\mathbf{X}_{Video}; \theta_v, \delta_f, \text{ReLU}), \tag{9}$$

Where $\mathbf{E}_{Audio}$, $\mathbf{E}_{Frame}$, $\mathbf{E}_{Pose}$, and $\mathbf{E}_{Video}$ represents the output of the convolutional layers for Audio, Frame, Pose, and Video features respectively.

Component-wise attention is applied to each feature, followed by a softmax function φ. Let $\mathbf{Q}_{Audio}$, $\mathbf{K}_{Audio}$, and $\mathbf{V}_{Audio}$ are the query, key, and value matrices computed from the input sequence $\mathbf{E}_{Audio}$, embedded into appropriate matrices. The component-wise attention and the final feature representation are computed as follows:

$$\alpha_{Audio} = \varphi\left(\frac{Q_{Audio}K^T_{Audio}}{\sqrt{d_k}}\right)V_{Audio}, \tag{10}$$

where α_{Audio} will be a weighted sum of the corresponding elements in $\mathbf{V}_{Audio}$, where the weights are determined by the attention scores computed using $\mathbf{Q}_{Audio}$ and $\mathbf{K}_{Audio}$. Similar formulations apply to Frame, Pose, and Video features to obtain the α_{Frame}, α_{Pose}, and α_{Video}.

For each feature modality, the output from the corresponding attention module (α_{Audio}, α_{Frame}, α_{Pose}, α_{Video},) undergoes processing through a dense layer followed by a softmax layer. This sequential operation results in the class probabilities for three distinct emotion classes (see Figure 2). Let $\mathcal{D}(\cdot; \lambda)$ denote the dense module with parameters λ. The output is given by:

$$A_{prob} = \varphi(\mathcal{D}(\alpha_{\text{Audio}}; \lambda_a)), \tag{11}$$

$$F_{prob} = \varphi(\mathcal{D}(\alpha_{\text{Frame}}; \lambda_f)), \tag{12}$$

$$P_{prob} = \varphi(\mathcal{D}(\alpha_{\text{Pose}}; \lambda_p)), \tag{13}$$

$$V_{prob} = \varphi(\mathcal{D}(\alpha_{\text{Video}}; \lambda_v)), \tag{14}$$

Where φ is a softmax function used for multiclass classification problems. The model incorporates a decision fusion mechanism to weigh the importance of each input modality. Finally, we have applied the decision fusion method, which involves the aggregation of the output scores produced by the softmax layers of each feature modality through a weighted sum operation in the following manner:

$$\hat{\mathbf{Y}} = \gamma_1 A_{prob} + \gamma_2 F_{prob} + \gamma_3 P_{prob} + \gamma_4 V_{prob} \tag{15}$$

where A_{prob}, F_{prob}, P_{prob}, and V_{prob} are the predicted probabilities for three classes using four different modalities (i.e. Audio, Frame, Pose, and Video), and γ_1, γ_2, γ_3, and γ_4 are the decision weights where each $\gamma_i \in [0, 1]$ and $\sum_i \gamma_i = 1$. And $\hat{\mathbf{Y}}$ represents the predicted class probability. We perform a grid search with a step size of 0.05 to identify the optimal weights(γ_1, γ_2, γ_3, and γ_4), maximizing classification accuracy. Results are reported for best-performing weight values.

Bimodal / Trimodal: For bimodal and trimodal decision fusion, unimodal descriptors are combined using weighted sums, employing equations similar to those used for multimodal decision fusion (as described above). Each feature modality undergoes processing through a single 1D-CNN layer, followed by an attention layer and, subsequently a softmax layer, resulting in class probabilities for each feature cue. The resultant outputs are then merged using the weighted sum methodology to produce the final class probabilities.

Unimodal: Motivated by the advancements in monomodal architectures [30], we have attempted to evaluate the predictive efficacy of unimodal methodologies. To achieve this, we deployed two distinct architectures: a 1D-CNN capable of capturing local data patterns and an Attention-based architecture proficient in capturing both local and global dependencies within the input sequence. The 1D-CNN model includes an input layer, one 1D-CNN layer, a dropout layer (dropout rate of 0.2), ReLU activation, and an output layer with 3 neurons with softmax activation. Conversely, the Attention-based architecture consists of an input layer, an attention layer, and an output layer with 3 neurons and softmax activation. These models are trained on features extracted from each individual modality.

Early Fusion (EF): For ablation purpose, we have also explored fusing features from each modality using feature fusion before forwarding them to any softmax (output) layer. Early fusion combines features from different modalities before they undergo further processing in subsequent neural network layers. This method enables the model to learn from multiple information sources simultaneously, potentially enhancing its understanding and performance.

4 Experiments and Results

The experimental details, ablation studies, and discussion of the results are described in this section.

4.1 Dataset

To evaluate the performance of proposed approach, we employ a publicly available VGAF dataset [31,32] for group emotion classification. The VGAF (Video Group Affect) dataset comprises YouTube videos featuring a diverse range of genders, ethnicity, event types, group sizes and poses. VGAF dataset has three types of group-level emotions: positive, negative, and neutral. The dataset contains a total of 326 videos and a total samples of 4,183; in these samples, 2,661 are train samples, 766 are validation samples, and 756 are test samples respectively. The videos display varying resolutions, and 5-second labeled clips are extracted from the entire video, with frame rates ranging from 13 to 30 frames per second. Within the VGAF dataset, each 5-second video clip sourced from the same origin might be annotated with distinct emotional labels, resulting in emotional ambiguity that complicates recognition.

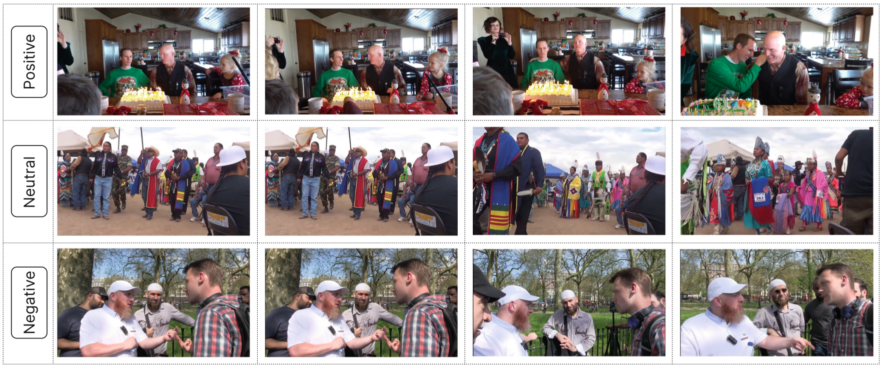


Fig. 3. Sample video frames from three different video samples containing positive, neutral, and negative classes from the VGAF dataset.

Table 1. Details of the parameters tuned using KerasTuner along with their potential choices and final value. Here CCE represents Categorical cross-entropy.

Parameter detail	Options	Selected Value
Learning rate	{1e−2, 1e−3, 1e−4, 1e−5}	1e−4
Validation split	{10%, 15%, 20%}	10%
Batch Size	{64, 100, 128, 256}	128
No. of filters (Conv1D layer)	{128, 256, 512, 1024}	256
Kernel size (Conv1D layer)	{3, 5, 7, 9}	5
Network Optimizer	{Adam, RMSprop, Adagrad, SGD}	Adam
Loss Function	{Categorical cross-entropy, Sparse Categorical cross-entropy }	CCE
Activation Function	{ReLU, Sigmoid, Tanh, SeLU}	ReLU
Epochs	{100, 150, 200, 250, 300, 350}	250

4.2 Training strategy and parameter tuning

We have trained our model using NVIDIA QUADRO P5000 GPU with 16 GB GDDR5X GPU memory, 2560 CUDA cores, and 16 GB Virtual RAM with Ubuntu 22.04 OS machine. The testing of the trained network has been carried out using 12th Gen Intel(R) Core i7-12650H, 2300 Mhz 16 GB RAM CPU machine with Windows Operating System.

We have used KerasTuner [26] to find out the best parameters for our approach. Table 1 contains the information related to the choice of parameters and final values chosen using parameter tuning.

4.3 Prediction Settings

We are utilizing the VGAF dataset (public) to test the performance of the proposed model. The VGAF dataset includes distinct samples for training, validation, and testing. However, as the test set is not publicly accessible, we have only obtained the training and validation sets from the dataset owners. For experimental purposes, we have trained the network using the training set and tested it using the validation set. It's important to note that the validation set has not been exposed to the model during training, serving as an unseen split for testing purposes. Our models are fine-tuned using the validation split, which comprises 10% of the training data and do not rely on external datasets for fine-tuning or pretraining. The validation split is decided using KerasTuner (see Table 1). We train the model for 250 epochs with early stopping. Results are reported in terms of classification accuracy.

Table 2. Comparison of overall accuracy (%) with state-of-the-art methods on the validation set of the VGAF Dataset. A: Audio, V: Video, H: Holistic, F: Face, L: Language, SD: Synthetic Data, P: Pose, Fr: Frame, Val: Validation, Acc: Accuracy.

Reference	**Features**	**Method**	**Val Acc**
Dhall et al. [10]	A,V	Inception-V3, CNN-LSTM	51.30
Petrova et al. [28]	SD	VGG-19	52.36
Sharma et al. [31]	A, H, F	Early fusion with LSTM and MLP	61.61
Pinto et al. [29]	A, V	Resnet-50, BiLSTM, and fusion SVM	62.40
Wang et al. [38]	A, V, L	K-injection network	66.19
Augusma et al. [2]	A, V, SD	ViT, Transformer Encoder, Cross Attention	78.72
Proposed	A, V, P, Fr	Multimodal decision fusion	**81.98**

4.4 Results & Discussion

This section elaborates on the results produced with the proposed approach. A summary of the results with different combinations of fusion methods and classification models is present in Table 4. A summary of the different combinations of features with the proposed approach is available in Table 3. While, Table 2 contains the comparison of the results from the proposed approach with the state-of-the-art(SoTA) approaches.

Overall Results We present the overall and class-wise classification results on the validation set in Table 3. Our proposed multimodal fusion method outperform baseline (see Table 2) accuracy by 29.88%. We make following observations:

- Fusion approaches outperform their unimodal counterpart and showing the efficacy of combining information from multiple sources.

Table 3. Classwise and overall accuracy (%) of proposed model across various feature combinations on validation set.

Features used	**Classwise(Val: Acc)**			**Overall**
	Positive	Neutral	Negative	(Val: Acc)
Audio Only	41.67	88.17	64.22	64.09
Frame Only	60.24	72.43	93.21	72.06
Pose Only	27.53	60.43	81.34	48.30
Video Only	73.65	64.25	50.12	60.57
Audio + Frame	58.23	82.43	95.33	75.84
Audio + Pose	37.56	91.65	82.23	67.62
Audio + Video	53.17	88.76	77.12	71.40
Frame + Pose	59.56	72.41	93.46	72.06
Frame + Video	61.34	74.49	92.34	73.10
Pose + Video	44.89	65.87	80.76	61.09
Audio + Frame + Pose	57.69	82.11	84.93	75.84
Audio + Frame + Video	62.34	83.34	95.55	77.41
Frame + Pose + Video	61.25	74.55	92.31	73.10
Audio + Pose + Video	43.67	90.92	91.32	71.93
Audio + Pose + Frame + Video	66.83	84.78	94.16	81.98

- Proposed approach has resulted in 81.98% accuracy, showcasing the efficacy of the proposed approach.
- Trimodal Fusion ($Audio + Frame + Video$) is also resulted in comparable performance.

Feature-Affect Class Correlation Analysis In order to discern the relationship between the utilized features and affect classes (positive, neutral, or negative), we trained unimodal features and computed the accuracies for each class (refer to Table 3). Notably, we observed that:

- Audio exhibits strong predictive capability for the neutral class. This phenomenon may be attributed to the fact that audio contains contextual cues beyond verbal content, with vocal expressions commonly associated with neutral emotions—such as calm and composed tone and clearer vocal intonations are predominantly conveyed through audio signals.
- Frame features, which capture the spatio-temporal context, and Pose Features, which contain finer movements of individuals, demonstrate superior performance in predicting negative affect. Negative emotions tend to display significant deviations in facial expressions and body postures across consecutive frames of a video sequence compared to the other two classes. Pose and frame-level features serve as baseline descriptors of negative emotional states by establishing a reference point against which deviations are detected. In

Table 4. Results of the ablation study investigating a multimodal model incorporating audio, pose, frame, and video data over the validation data set. Accuracy (Acc) is reported in percentages (%), with two fusion methods evaluated: Early Fusion (EF) and Decision Fusion (DF). Here, κ is used for the Number of filters in the 1D-CNN layer. Γ is used to denote the fusion method. A number in a bracket denotes the number of layers used.

Model	Γ	κ	**Classwise(Val: Acc)**			**Overall**
			Positive	Neutral	Negative	**(Val: Acc)**
1D-CNN (1)	EF	256	72.24	54.89	67.15	64.62
1D-CNN (1)	DF	256	49.44	78.34	69.78	65.01
Attention (1)	EF	-	69.23	56.23	70.54	65.01
Attention (1)	DF	-	45.14	60.78	71.33	56.91
1D-CNN (1) + Attention (1)	EF	256	64.83	84.78	94.16	78.72
LSTM (1)	EF	-	55.62	60.35	73.36	61.61
LSTM (1)	DF	-	32.11	77.14	76.08	59.13
LSTM (1)+ Attention (1)	EF	-	75.43	65.34	82.93	69.76
LSTM (1)+ Attention (1)	DF	-	45.54	79.64	88.91	65.18
1D-CNN (2)	DF	256	54.63	73.92	71.73	65.79
Attention (2)	DF	-	26.15	86.78	70.10	58.87
1D-CNN (2) + Attention (1)	DF	256	67.54	91.07	93.47	80.18
1D-CNN (2)+ Attention (1)	DF	128	63.57	93.57	87.50	79.28
1D-CNN (2)+ Attention (2)	DF	256	35.23	80.00	82.65	59.53
1D-CNN (3)+ Attention (2)	DF	256	37.20	79.00	81.65	59.34
1D-CNN (3) + Attention (1)	DF	256	68.87	94.28	76.08	79.89

contrast, positive and neutral emotions may exhibit less pronounced deviations from this baseline, highlighting the relevance of features that capture intensity or arousal.
- TimeSformer-based features (Video only) demonstrate their efficacy in predicting the "Positive" class. Positive emotions typically entail dynamic temporal changes, such as excitement or happiness. TimeSformer models excel at capturing temporal dependencies and long-range interactions within sequential data, allowing them to encode the temporal dynamics associated with positive emotions effectively. Furthermore, video features yield comparable results for the "Negative" class, indicating their capability to preserve dynamic temporal changes.

Comparison with state-of-the-art(SoTA) approaches The outcomes of our proposed method on the VGAF dataset have been presented and juxtaposed with existing methodologies in Table 2. Our approach exhibits superior performance compared to benchmark methods. Augusma et al. [2], utilizing two-stream features (A, V) and Synthetic Data, achieved 78.72% accuracy, while the baseline by Dhall et al. [10] was 52.10%. We demonstrate that capturing small details, such as pose position, context, and surrounding objects, contributes to better modeling of group emotions.

Table 5. Results of the ablation study investigating an unimodal model incorporating audio, pose, frame, and video data over the validation data set. Accuracy (Acc) is reported in percentages (%). A number in a bracket denotes the number of layers used.

Model	**Feature**	**Classwise(Val: Acc)**			**Overall**
		Positive	Neutral	Negative	**(Val: Acc)**
1D-CNN (1)	Audio	30.79	82.14	48.91	53.91
1D-CNN (1)	Frame	42.38	57.85	74.44	55.74
1D-CNN (1)	Pose	12.10	62.85	78.26	45.43
1D-CNN (1)	Video	72.28	66.78	54.63	63.31
Attention (1)	Audio	24.17	88.57	49.45	53.78
Attention (1)	Frame	15.33	67.60	19.12	36.55
Attention (1)	Pose	17.54	71.50	17.23	38.90
Attention (1)	Video	72.82	61.78	26.82	50.65
1D-CNN (2)	Audio	35.09	74.28	53.80	53.91
1D-CNN (2)	Frame	43.70	54.28	71.19	54.17
1D-CNN (2)	Pose	11.34	60.35	80.84	44.12
1D-CNN (2)	Video	71.19	70.00	44.37	60.18

Ablation Study Additionally, we conducted an ablation study using various fusion strategies (early fusion and late fusion) with three architectures (CNN, LSTM, and Attention) across four features: audio, frame, pose, and video. The results are depicted in Table 4, along with two different unimodal architectures for each feature input shown in Table 5. Table 4 demonstrates that certain classes, such as Neutral (EF (1D CNN + Attention)), perform exceptionally well across different configurations. In Table 5, some features, like video, contribute significantly, while others, like pose, show negligible performance with certain architectures. This study helped us determine the optimal architecture and fusion strategy for classifying video-group emotions.

5 Conclusion and future work

In this paper, we introduce a multimodal decision fusion model designed to predict group affect at the video level. We integrate various modalities, including audio, video, and pose, into multimodal fusion settings and assess their performance. Our experimental findings reveal that the audio modality exhibits a strong predictive capacity for negative emotions, attributed to its inclusion of contextual cues beyond verbal content. Additionally, the pose modality effectively captures consistent facial expressions and subtle body movements characteristic of the neutral class. Multimodal fusion surpasses unimodal, bimodal, and trimodal fusion approaches, showcasing its ability to retain more information from parallel pipelines. Experimental results demonstrate superior performance compared to state-of-the-art approaches. Moreover, our analysis extends beyond multiclass emotion classification to explore correlation patterns between classes and features (interpretability). We demonstrate that features extracted from audio, pose, frame, and video-level modalities contribute to video-based group affect recognition in real-world settings.

One limitation of our study is the lack of additional datasets available to assess the generalizability of the proposed model. For future research, we propose to enhance the model's capabilities by integrating more advanced models for feature extraction and employing new techniques for fusing information from multi-branch sources. Additionally, we plan to evaluate the performance of our approach on diverse datasets with varying labels, such as sarcasm detection [7], to further validate its effectiveness. Furthermore, we aim to prioritize improving the model's generalizability to ensure its applicability across different contexts and scenarios.

References

1. Abdu, S.A., Yousef, A.H., Salem, A.: Multimodal video sentiment analysis using deep learning approaches, a survey. Information Fusion **76**, 204–226 (2021)
2. Augusma, A., Vaufreydaz, D., Letué, F.: Multimodal group emotion recognition in-the-wild using privacy-compliant features. In: Proceedings of the 25th International Conference on Multimodal Interaction. pp. 750–754 (2023)
3. Baevski, A., Zhou, Y., Mohamed, A., Auli, M.: wav2vec 2.0: A framework for self-supervised learning of speech representations. Advances in neural information processing systems **33**, 12449–12460 (2020)
4. Balaji, B., Oruganti, V.R.M.: Multi-level feature fusion for group-level emotion recognition. In: Proceedings of the 19th ACM international conference on multimodal interaction. pp. 583–586 (2017)
5. Belova, N.S.: Group-level affect recognition in video using deviation of frame features. In: Analysis of Images, Social Networks and Texts: 10th International Conference, AIST 2021, Tbilisi, Georgia, December 16–18, 2021, Revised Selected Papers. vol. 13217, p. 199. Springer Nature (2022)
6. Bertasius, G., Wang, H., Torresani, L.: Is space-time attention all you need for video understanding? In: Proceedings of the International Conference on Machine Learning (ICML) (July 2021)
7. Castro, S., Hazarika, D., Pérez-Rosas, V., Zimmermann, R., Mihalcea, R., Poria, S.: Towards multimodal sarcasm detection (an _obviously_ perfect paper). arXiv preprint arXiv:1906.01815 (2019)
8. Collins, J.A., Olson, I.R.: Knowledge is power: How conceptual knowledge transforms visual cognition. Psychonomic bulletin & review **21**, 843–860 (2014)

9. Constantin, M.G., Ştefan, L.D., Ionescu, B., Demarty, C.H., Sjöberg, M., Schedl, M., Gravier, G.: Affect in multimedia: Benchmarking violent scenes detection. IEEE Trans. Affect. Comput. **13**(1), 347–366 (2020)
10. Dhall, A., Sharma, G., Goecke, R., Gedeon, T.: Emotiw 2020: Driver gaze, group emotion, student engagement and physiological signal based challenges. In: Proceedings of the 2020 International Conference on Multimodal Interaction. pp. 784–789 (2020)
11. Evtodienko, L.: Multimodal end-to-end group emotion recognition using cross-modal attention. arXiv preprint arXiv:2111.05890 (2021)
12. Feichtenhofer, C., Fan, H., Malik, J., He, K.: Slowfast networks for video recognition. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 6202–6211 (2019)
13. Ferreira, P.M., Marques, F., Cardoso, J.S., Rebelo, A.: Physiological inspired deep neural networks for emotion recognition. IEEE Access **6**, 53930–53943 (2018)
14. Guo, X., Zhu, B., Polanía, L.F., Boncelet, C., Barner, K.E.: Group-level emotion recognition using hybrid deep models based on faces, scenes, skeletons and visual attentions. In: Proceedings of the 20th ACM International Conference on Multimodal Interaction. pp. 635–639 (2018)
15. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
16. Howse, J.: OpenCV computer vision with python, vol. 27. Packt Publishing Birmingham, UK (2013)
17. Huang, X., Dhall, A., Goecke, R., Pietikäinen, M., Zhao, G.: Analyzing group-level emotion with global alignment kernel based approach. IEEE Trans. Affect. Comput. **13**(2), 713–728 (2019)
18. Jin, B.T., Abdelrahman, L., Chen, C.K., Khanzada, A.: Fusical: Multimodal fusion for video sentiment. In: Proceedings of the 2020 International Conference on Multimodal Interaction. pp. 798–806 (2020)
19. Jocher, G., Chaurasia, A., Qiu, J.: Ultralytics YOLO (Jan 2023), https://github.com/ultralytics/ultralytics
20. Kelly, J.R., Barsade, S.G.: Mood and emotions in small groups and work teams. Organ. Behav. Hum. Decis. Process. **86**(1), 99–130 (2001)
21. Li, S., Deng, W.: Deep facial expression recognition: A survey. IEEE Trans. Affect. Comput. **13**(3), 1195–1215 (2020)
22. Liu, C., Jiang, W., Wang, M., Tang, T.: Group level audio-video emotion recognition using hybrid networks. In: Proceedings of the 2020 International Conference on Multimodal Interaction. pp. 807–812 (2020)
23. Magnani, L., Civita, S., Massara, G.P.: Visual cognition and cognitive modeling. Human and machine vision: Analogies and divergencies pp. 229–243 (1994)
24. Morris, R.G., Tarassenko, L., Kenward, M.: Cognitive systems-Information processing meets brain science. Elsevier (2005)
25. Niedenthal, P.M., Brauer, M.: Social functionality of human emotion. Annu. Rev. Psychol. **63**, 259–285 (2012)
26. O'Malley, T., Bursztein, E., Long, J., Chollet, F., Jin, H., Invernizzi, L., et al.: Kerastuner. https://github.com/keras-team/keras-tuner (2019)
27. Pan, C., Yu, D., Sijiang, L., Zhen, G., Lei, Y.: Group emotion recognition based on multilayer hybrid network. In: 2018 IEEE 3rd International Conference on Image, Vision and Computing (ICIVC). pp. 173–177. IEEE (2018)

28. Petrova, A., Vaufreydaz, D., Dessus, P.: Group-level emotion recognition using a unimodal privacy-safe non-individual approach. In: Proceedings of the 2020 International Conference on Multimodal Interaction. pp. 813–820 (2020)
29. Pinto, J.R., Gonçalves, T., Pinto, C., Sanhudo, L., Fonseca, J., Gonçalves, F., Carvalho, P., Cardoso, J.S.: Audiovisual classification of group emotion valence using activity recognition networks. In: 2020 IEEE 4th International Conference on Image Processing, Applications and Systems (IPAS). pp. 114–119. IEEE (2020)
30. Savchenko, A.V., Makarov, I.: Neural network model for video-based analysis of student's emotions in e-learning. Optical Memory and Neural Networks **31**(3), 237–244 (2022)
31. Sharma, G., Dhall, A., Cai, J.: Audio-visual automatic group affect analysis. IEEE Trans. Affect. Comput. **14**(2), 1056–1069 (2021)
32. Sharma, G., Ghosh, S., Dhall, A.: Automatic group level affect and cohesion prediction in videos. In: 2019 8th International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW). pp. 161–167. IEEE (2019)
33. Tian, Y., Lu, G., Yan, Y., Zhai, G., Chen, L., Gao, Z.: A coding framework and benchmark towards low-bitrate video understanding. IEEE Transactions on Pattern Analysis and Machine Intelligence (2024)
34. Tian, Y., Lu, G., Zhai, G., Gao, Z.: Non-semantics suppressed mask learning for unsupervised video semantic compression. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 13610–13622 (2023)
35. Tian, Y., Yan, Y., Zhai, G., Guo, G., Gao, Z.: Ean: event adaptive network for enhanced action recognition. Int. J. Comput. Vision **130**(10), 2453–2471 (2022)
36. Veltmeijer, E.A., Gerritsen, C., Hindriks, K.V.: Automatic emotion recognition for groups: a review. IEEE Trans. Affect. Comput. **14**(1), 89–107 (2021)
37. Wang, Y., Song, W., Tao, W., Liotta, A., Yang, D., Li, X., Gao, S., Sun, Y., Ge, W., Zhang, W., et al.: A systematic review on affective computing: Emotion models, databases, and recent advances. Information Fusion **83**, 19–52 (2022)
38. Wang, Y., Wu, J., Heracleous, P., Wada, S., Kimura, R., Kurihara, S.: Implicit knowledge injectable cross attention audiovisual model for group emotion recognition. In: Proceedings of the 2020 international conference on multimodal interaction. pp. 827–834 (2020)
39. Xiao, F., Lee, Y.J., Grauman, K., Malik, J., Feichtenhofer, C.: Audiovisual slowfast networks for video recognition. arXiv preprint arXiv:2001.08740 (2020)
40. Zadeh, A., Chan, M., Liang, P.P., Tong, E., Morency, L.P.: Social-iq: A question answering benchmark for artificial social intelligence. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8807–8817 (2019)
41. Zhang, K., Li, Y., Wang, J., Cambria, E., Li, X.: Real-time video emotion recognition based on reinforcement learning and domain knowledge. IEEE Trans. Circuits Syst. Video Technol. **32**(3), 1034–1047 (2021)
42. Zhao, S., Ma, Y., Gu, Y., Yang, J., Xing, T., Xu, P., Hu, R., Chai, H., Keutzer, K.: An end-to-end visual-audio attention network for emotion recognition in user-generated videos. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 34, pp. 303–311 (2020)

CSSR: Cross-and Self-feature Transformer with High-Frequency Feature Alignment for Reference-Based Super-Resolution

Seonggwan Ko[1,2] and Donghyeon Cho[3](✉)

[1] Chungnam National University, Daejeon, South Korea
[2] HL Klemove, Seongnam, South Korea
[3] Department of Computer Science, Hanyang University, Seoul, South Korea
doncho@hanyang.ac.kr

Abstract. Reference-based super-resolution (RefSR) utilizes an external high-resolution reference (Ref) image to transfer detailed textures to a low-resolution (LR) image, resulting in improved performance over single-image super-resolution (SISR) methods. The main challenge in RefSR is to find correspondences between the LR and Ref images, and accurately convey the rich texture information of the Ref image. However, this becomes difficult when the similarity between the LR and Ref images is low or there is ambiguity in the matching stage. To address these challenges, we propose a novel cross-and self-feature transformer (CSFT) which integrates not only the rich visual features of the Ref image, but also the internal information within the input LR image. In addition, we introduce a high-frequency feature alignment (HFFA) module to robustly fuse the features of the LR and Ref images even in areas where alignment is ambiguous. Based on the proposed CSFT and HFFA modules, we define a new RefSR pipeline, referred to as CSSR, where each module is structured with multi-scales. The CSSR can fully utilize textural information in both Ref and LR images and achieve outstanding performance, even when feature matching between Ref and LR images is challenging. Various experiments have been conducted to verify the effectiveness of CSSR, both quantitatively and qualitatively. The source codes is available at https://github.com/SeonggwanKo/CSSR.

Keywords: Reference-based super-resolution · Transformer · deep neural networks

1 Introduction

Single image super-resolution (SISR) [6,16,19,32,43] is an extensively studied field that aims to generate a high-resolution (HR) image from a single low-resolution (LR) image. Due to the loss of high-frequency details during the

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/978-3-031-78305-0_27.

A. Antonacopoulos et al. (Eds.): ICPR 2024, LNCS 15321, pp. 419–434, 2025.
https://doi.org/10.1007/978-3-031-78305-0_27

down-sampling process, the ground truth (GT) HR image corresponding to an input LR image is not unique. Therefore, SISR is a highly ill-posed problem characterized by a complex one-to-many relationship. To tackle this problem, there have been various attempts, such as utilizing natural image priors (i.e., gradient profile [27]) or a generative adversarial network (GAN) [10]. Even though these approaches can produce visually plausible images, the output may deviate from the original HR image.

Recently, to mitigate this issue, reference-based super-resolution (RefSR) that utilizes additional HR images has been introduced [14,21,34,36,44,45]. Most RefSR methods restore a high-quality HR image by accurately finding the correspondence between the input LR and Ref images and then transferring the detailed information of the Ref image. However, there are still challenging issues that need to be addressed in order to recover detailed textures. First, the Ref image becomes less effective in generating the final SR result when there is a low similarity between the input LR and Ref images. This can be seen in the top of Figure 1, where the performance of DATSR [2], which is one of the recent RefSR methods, drops significantly under such circumstances. To restore the red box in the LR image, it may be more beneficial to obtain useful information from other scales of the LR image (*i.e.*, the yellow box of the top in Fig. 1) rather than the low-similarity Ref image. Second, even slight variations in viewpoint or illumination between Ref and LR images can make feature alignment ambiguous, and degrade performance. As shown in the bottom of Fig. 1, it seems simple to restore the red box in the input by using textures in the yellow box in the Ref image, but DATSR suffers from difficulties in feature alignment due to ambiguity caused by aforementioned variations. In this case, it can be helpful to

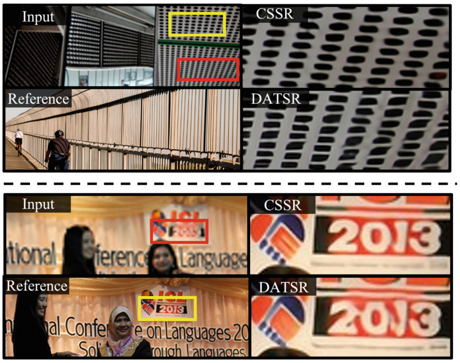


Fig. 1. (Top) The dissimilarities between the two images prevent successful matching, resulting in an alignment performance drop. (Bottom) Although the two images share similarities, their different viewpoints and brightness prevent leads in misalignment. Our approach of utilizing cross-and self-features transformer (CSFT) and high-frequency feature alignment (HFFA) in both LR and Ref pairs is effective in mitigating these issues.

perform feature alignment based on the high-frequency components excluding the low-frequency components.

In this paper, we effectively resolve two aforementioned issues as follows. To address the problem of low similarity between two images, we introduce a novel cross-and self-feature transformer (CSFT), a module that uses detailed textures from the Ref image as well as internal information within the input LR image. Specifically, the CSFT module comprises two cross-feature transformers (CFTs) and a self-feature transformer (SFT). The first CFT aggregates high-frequency features extracted from the Ref image, whereas the second CFT combines restored LR features utilizing a pre-trained SISR model. Subsequently, the SFT enhances features even more by capitalizing on internal self-similarity information. Meanwhile, to overcome ambiguity problems by the low-frequency feature components, we propose a high-frequency feature alignment (HFFA) module to align textures in the Ref features by effectively conveying residual features. The proposed HFFA performs better feature alignment in detailed textured areas compared to the existing flow-based deformable convolution network [13,14]. Based on the proposed CSFT and HFFA modules, we develop CSSR, a new RefSR method that is robust against low similarity with the Ref image or ambiguity in texture regions during feature alignment. We summarize our key contributions as follows:

- We propose a novel cross-and self-feature transformer (CSFT) for RefSR, which effectively utilizes internal information in the input LR image, as well as the cross-information between Ref and LR images.
- To align detailed information in the Ref image to the grid of the input LR efficiently, we suggest a high-frequency feature alignment (HFFA) module.
- Through extensive experiments on benchmarks, we demonstrate that our model achieves notably high performance even when input and Ref images are less similar and there are ambiguities in the feature alignment.

2 Related Works

2.1 Single Image Super-Resolution

Single image super-resolution (SISR) is a task to recover the high-resolution (HR) image from the low-resolution (LR) image. Most of the deep learning-based SISR methods are built on the convolutional neural network (CNN), and have achieved much better performance than conventional algorithms [3,9,15, 30,37,38]. As a pioneering work, [6] proposed an SRCNN consisting of three convolutional layers for SISR. Based on SRCNN, there are various efforts to design very deep networks with residual connections in [18–20,22,29,43]. After that, to further boost up the performance by exploiting long-range dependencies within the image, the attention mechanisms have been adopted in [4,5,17,24, 42,46]. However, despite the efforts aforementioned, most existing SISR models tend to produce smoothed results. To overcome this issue, [16] designed SRGAN using the adversarial loss to generate a realistic HR image. Later, [32] proposed

ESRGAN exploiting relativistic discriminator to reduce artifacts in SRGAN and produce more realistic results.

2.2 Reference-Based Super-Resolution

Unlike SISR, reference-based super-resolution (RefSR) can generate a realistic and detailed HR image by using an additional HR Ref image. The main issue of RefSR is to find corresponding patches between the LR and Ref images, and transfer features of the Ref image. Recently, CNN-based RefSR models have been suggested to utilize high-frequency information from the Ref image. [45] proposed CrossNet, which transfers textures in the Ref image by warping features of the Ref image. However, the performance is degraded due to inaccurate alignment. Then, SRNTT [44] improved matching performance in a patch-match-based correlation method. Although the matching accuracy has been improved, SRNTT has a limitation in delivering detailed textures of the Ref image because it relies on VGG19 [26] that focuses on semantic features for the classification. To align the feature more accurately, [36] proposed a learnable texture transformer while [25] utilized a deformable convolution. [21] introduced MASA, which is based on the coarse and fine matching module to reduce the computational cost of finding correspondences between the LR and Ref images. Furthermore, [34] applied the coarse-to-fine patch-match module, which requires lower resources. To enhance the feature extractor against rotation and scale variations, C^2-matching was proposed by [14] based on contrastive learning and knowledge distillation in patch matching steps. Recently, [13] proposed a method that reduces the reference mis-use and under-use issues by decoupling the texture transfer module and the SR module. [2] proposed DATSR that improves performance via deformable attention with multiple reference images. Moreover, [41] introduced RRSR that adopts progressive alignment with multiple reference images. [40] proposed LMR that introduces an efficient alignment for multiple references. Most previous studies have focused on improving matching accuracy and alignment to effectively convey rich visual features from the Ref image. However, in cases where similar information is lacking in the Ref image, there is a tendency for performance to degrade. Therefore, this paper not only utilizes Ref features and the SISR model as in [13], but also proposes a cross- and self-feature transformer (CSFT) module that comprehensively leverages the internal information inherent in the input LR image. Furthermore, to address ambiguity issues during the feature alignment that can arise in textured regions, a high-frequency feature alignment (HFFA) module is introduced. By seamlessly integrating CSFT and HFFA, state-of-the-art performance is achieved on the RefSR benchmarks.

3 Proposed Method

The proposed CSSR network is implemented in the manner of multi-scale. For the sake of simplicity, we describe the process on a single scale. An overview

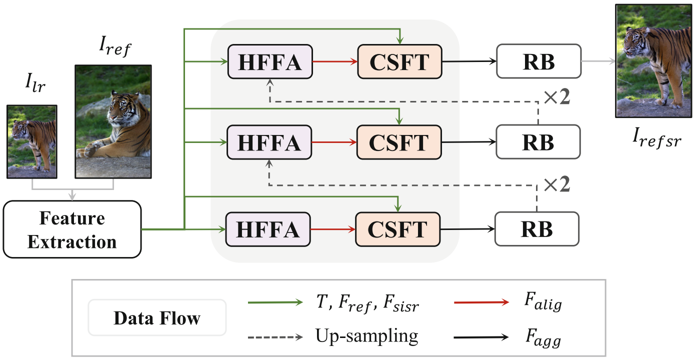


Fig. 2. Overview of CSSR network. Our CSSR consists of high-frequency feature alignment (HFFA), cross-and self-feature transformer (CSFT) and residual blocks (RB) with a multi-scale structure.

of CSSR network is shown in Figure 2. We first extract the features of the LR and Ref images for matching to obtain a warping map, and then align the features of the Ref image through the high-frequency feature alignment (HFFA) module. Next, the cross-and self-feature transformer (CSFT) module aggregates the features of the Ref and LR images with cross-and self-information to generate more detailed features. Finally, the decoder generates the final output.

3.1 Feature Matching

To find correspondences between the feature of the Ref image I_{ref} and the feature of the LR image I_{lr}, the correlation matrix between features of I_{ref} and I_{lr} is typically used in RefSR tasks. In our CSSR, similar to [13,45], we employ an upsampled image I_{sisr} from using the SISR model [32] instead of I_{lr}. To be specific, we train a shared feature extractor similar to a method as introduced in [14], but utilize SISR-Ref pairs to reduce the geometric transformation gap. Then, we extract features of I_{sisr} and I_{ref} through this trained shared feature extractor. After that, we compute the correlation matrix M between two features of I_{sisr} and I_{ref}. Note that these features are only used to compute M. With this correlation matrix, we compute a warping map T_i that transforms pixels of I_{ref} for the position i of the I_{sisr} as

$$T_i = \underset{j}{\operatorname{argmax}}\, M_{i,j}. \tag{1}$$

In contrast to [13], we use multi-scale flow T for every step. Meanwhile, since the feature to be used for the final super-resolved output should have rich visual information of I_{ref} and I_{sisr} rather than information useful for matching, we utilize multi-scale features F_{ref} extracted from a pretrained VGG19 [26] and an upsampled features F_{sisr} obtained by the SISR model. In summary, the estimated T, F_{ref} and F_{sisr} are passed to HFFA module. We discuss the feature extractor and each multi-scale feature in the supplementary material.

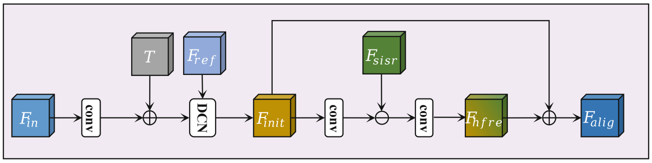


Fig. 3. The architecture of HFFA.

3.2 High-Frequency Feature Alignment

In this stage, we align F_{ref} to obtain detailed textures in the Ref image. The existing flow-based DCN struggles [13,14,34] to efficiently convey high-frequency information from I_{ref}, thus we propose high-frequency feature alignment (HFFA) inspired by [8]. Let F_{in} be the input feature for the proposed HFFA module at each scale step. Note that $F_{in} = F_{sisr}$ for the first step. For the rest of the steps, the output of the previous step is the input of the next step. (See the dot line in Fig. 2). First, we pass F_{in} to a convolutional layer, and add the warping map T estimated in (1) to obtain an offset O for stably training the DCN. We apply the DCN to F_{ref} with O to get the initially aligned features as

$$F_{init} = DCN(F_{ref}, O), \tag{2}$$

where $DCN(\cdot)$ means an operation of the DCN. Then, unlike previous methods [13,14,34], we compute a high-frequency residual features by subtraction from F_{sisr} as

$$F_{hfre} = conv(F_{sisr} - conv(F_{init})), \tag{3}$$

where $conv(\cdot)$ is a convolution operation. The final aligned features are computed by summing F_{hfre} and F_{init} as

$$F_{alig} = F_{hfre} + F_{init}. \tag{4}$$

Note that F_{hfre} contains the high-frequency details by subtracting F_{sisr}, whereas F_{init} has overall contents of visual information. Therefore, compared to previous methods, the proposed HFFA module can boost detailed texture in I_{ref}. The processes of HFFA module are shown in Figure 3.

3.3 Cross-and Self-Feature Transformer

Although promising results can already be obtained by using only HFFA module, in order to achieve further performance improvement, a module capable of more comprehensive aggregation of the information of the Ref image, the restored image from the SISR model, and the input LR image is required. Therefore, we newly introduce a novel cross-and self-feature transformer (CSFT) that is composed of two cross-feature transformers (CFTs) and a self-feature transformer (SFT). The first CFT module transfers rich texture features F_{ref} in the Ref

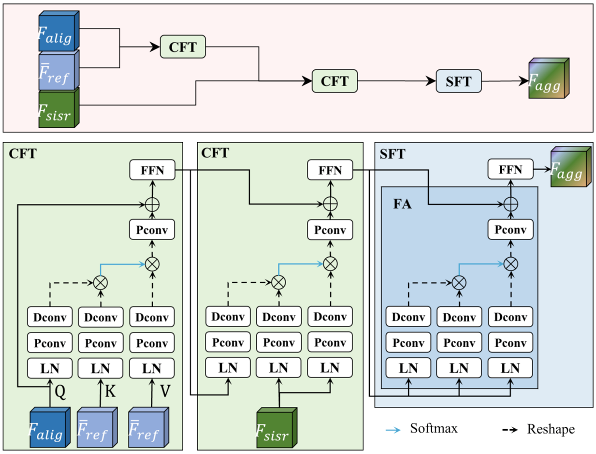


Fig. 4. The architecture of CSFT. CSFT module consists of two CFTs that enrich aligned features by transferring features from both the Ref and the restored images and a SFT that utilizes the internal information of aligned features for further improvement.

image to the aligned features F_{alig}, while the second CFT conveys restored features F_{sisr} from the pretrained SISR model. In other words, the two CFTs can aggregate both the complementary information of the Ref image and the restored image by the SISR model. Note that, to aggregate F_{ref} more easily, we initially warp F_{ref} into the grid of LR using T from Eq. (1) before passing it to the first CFT. We denote it as $\bar{F}_{ref}$. Then, the SFT is applied to further exploit internal similarity information within an image. Hence, a series of CFTs and the SFT is expressed as

$$F_{agg} = SFT(CFT(CFT(F_{alig}, \bar{F}_{ref}), F_{sisr})), \tag{5}$$

where F_{agg} is an aggregated feature by CSFT module while $CFT(\cdot)$ and $SFT(\cdot)$ are operations of CFT and SFT, respectively. As introduced in [39], query Q, key K and value V vectors are extracted by $Q = W_{dp}LN(f_1)$, $K = W_{dp}LN(f_2)$, $V = W_{dp}LN(f_3)$. Here, $LN(\cdot)$, W_{dp} denote a layer norm [1] and a series of a point-wise convolution and a depth-wise convolution. f_1, f_2 and f_3 are three input features. In more detail, both CFT and SFT consist of a feature attention and a feed-forward network, and they are defined as

$$FA(f_1, f_2, f_3) = W_p(V \cdot Softmax(K \cdot Q)) + f_1, \tag{6}$$
$$CFT(f_1, f_2) = FFN(FA(f_1, f_2, f_2)), \tag{7}$$
$$SFT(f_1) = FFN(FA(f_1, f_1, f_1)), \tag{8}$$

where $Softmax(\cdot)$, W_p, $FFN(\cdot)$ and $FA(\cdot)$ are a softmax operation, a point-wise convolution, the feed-forward network and the feature attention, respectively. We select GDFN [39] as the feed-forward network [7] to transfer high-quality texture information while suppressing less useful features. Our CFTs and SFT are illustrated in Figure 4. In the end, the aggregated feature F_{agg} in (5) is passed into the decoder to obtain the next step of the output feature. The output feature of decoders will be used as F_{in} for the next scale step. In the last scale step, the decoder generates the final restored image I_{refsr}. Note that decoders are composed of residual blocks.

3.4 Training Strategies

Training SISR model. To obtain a restored image I_{sisr}, we trained a SISR model from scratch using CUFED5 [44] dataset based on an ℓ_1-norm reconstruction loss. After training, all the parameters of the SISR networks are fixed. For a fair comparison, we adopt RRDB [32] applied in [13] as a SISR model.
Training CSSR. We follow all hyper-parameter setting and the training strategy following [14]. With fixed a pretrained SISR model and a feature extractor, our CSSR network is trained by a combination of an ℓ_1 norm reconstruction loss L_{rec}, a perceptual loss L_{per}, an adversarial loss L_{adv}. Therefore, I_{refsr} is compared to the GT image I_{gt} using L_{rec}, L_{per} and L_{adv} losses. Additionally, we adopts reciprocal loss L_{rtrr} [41] for better reconstruction. For computing L_{per}, we utilize features from $relu5_1$ layer in VGG19 [26]. Also, we adopt WGAN-GP [11] for L_{adv} without the pretrained model. The final loss for training is

$$L = \lambda_{rec}L_{rec} + \lambda_{per}L_{per} + \lambda_{adv}L_{adv} + \lambda_{rtrr}L_{rtrr}, \tag{9}$$

where λ_{rec},λ_{per},λ_{adv}, λ_{rtrr} are weights for each loss term, and set to 1.0, 10^{-4}, 10^{-6}, 0.6 respectively. We select ADAM optimizer with $\beta_1 = 0.9$ and $\beta_2 = 0.999$. The data augmentation contains random rotations, vertical and horizontal flips.

4 Experiments

4.1 Datasets and Evaluation

Our model is trained on CUFED5 [44] dataset containing 11,781 pairs of images, with each pair consisting of the HR and Ref images of size 160×160. For each pair, a LR image is generated from a HR image by downsampling. For evaluation, the test sets of CUFED5, SUN80 [28], Urban100 [12], Manga109 [23] and WRSR [14] datasets are used. Each dataset is composed as follows. In CUFED5, the testing set consists of 126 pairs of images, each pair consisting of a single input image with five Ref images that are another level of similarity. For testing, we only select a single Ref image that has the most similar level to the LR image. Sun80 provides 80 image pairs consisting of 20 Ref images corresponding to a single input image. WR-SR has 80 pairs of images, each pair containing a single image with a single Ref image. Urban100 and Manga109 only provide 100 and 109 images without Ref images. Thus, we randomly select HR images in the dataset as Ref images as conducted in other previous work [2]. The results of the RefSR methods are evaluated as PSNR and SSIM [33] with Y channel on YCbCr space. All experiments were performed with the scaling factor set to 4.

Table 1. Quantitative comparisons. We compare with SISR and RefSR models in terms of PSNR and SSIM. Note that CSSR utilizes only a *single* Ref image on CUFED5. The suffix '-rec' means reconstruction loss only version of RefSR models and without the suffix means full loss version. LMR* does not provide results from training with CUFED5, therefore we train LMR with CUFED5 for a fair comparison. The best results are marked **in bold**.

	Method	CUFED5		Sun80		Urban100		Manga109		WR-SR	
		PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
SISR	SRCNN [6]	25.33	0.745	28.26	0.781	24.41	0.738	27.12	0.850	27.27	0.767
	EDSR [19]	25.93	0.777	28.52	0.792	25.51	0.783	28.93	0.891	28.07	0.793
	RCAN [43]	26.06	0.769	29.86	0.810	25.42	0.768	29.38	0.895	28.25	0.799
	SRGAN [16]	24.40	0.702	26.76	0.725	24.07	0.729	25.12	0.802	26.21	0.728
	ESRGAN [32]	21.90	0.633	24.18	0.651	20.91	0.620	23.53	0.797	26.07	0.726
	SwinIR [17]	26.62	0.790	30.11	0.817	26.26	0.797	30.05	0.910	28.06	0.797
	SRFormer [46]	26.66	0.790	30.14	0.817	26.53	0.802	30.25	0.911	28.07	0.796
RefSR	SRNTT [44]	25.61	0.764	27.59	0.756	25.09	0.774	27.54	0.862	26.53	0.745
	SRNTT-rec [44]	26.24	0.784	28.54	0.793	25.50	0.783	28.95	0.885	27.59	0.780
	TTSR [36]	25.53	0.765	28.59	0.774	24.62	0.747	28.70	0.886	26.83	0.762
	TTSR-rec [36]	27.09	0.804	30.02	0.814	25.87	0.784	30.09	0.907	27.97	0.792
	DCSR [31]	25.39	0.733	-	-	-	-	-	-	-	-
	DCSR-rec [31]	27.30	0.807	-	-	-	-	-	-	-	-
	MASA [21]	24.92	0.729	27.12	0.708	23.78	0.712	27.34	0.849	-	-
	MASA-rec [21]	27.54	0.814	30.15	0.815	26.09	0.786	30.28	0.909	-	-
	C^2-Matching [14]	27.16	0.805	29.75	0.799	25.52	0.764	29.73	0.893	27.80	0.780
	C^2-Matching-rec [14]	28.24	0.841	30.18	0.817	26.03	0.785	30.47	0.911	28.32	0.801
	LMR* [40]	27.41	0.814	-	-	-	-	-	-	27.81	0.781
	LMR-rec* [40]	28.49	0.848	-	-	-	-	-	-	28.27	0.801
	AMSA [34]	27.31	0.803	29.83	0.803	25.60	0.770	29.79	0.896	-	-
	AMSA-rec [34]	28.50	0.849	30.29	0.819	26.18	0.789	30.57	0.914	-	-
	TDF [13]	27.37	0.816	28.85	0.768	25.80	0.776	30.12	0.889	27.40	0.769
	TDF-rec [13]	28.64	0.850	30.31	0.820	26.71	**0.807**	31.23	0.917	28.52	**0.807**
	DATSR [2]	27.95	0.835	29.77	0.800	25.92	0.775	29.75	0.893	27.87	0.787
	DATSR-rec [2]	28.72	0.856	30.20	0.818	26.52	0.798	30.49	0.912	28.34	0.805
	CSSR	28.30	0.840	29.95	0.806	26.02	0.782	30.01	0.897	28.29	0.797
	CSSR-rec	**29.03**	**0.858**	**30.36**	**0.821**	**26.87**	**0.807**	**31.37**	**0.918**	**28.55**	**0.807**

4.2 Evaluations

Quantitative evaluations. We compare the proposed CSSR with both existing SISR and RefSR methods. In particular, we adopt SRCNN [6], EDSR [19], RCAN [43], SRGAN [6], ESRGAN [32], SwinIR [17], SRFormer [46] as SISR methods, and SRNTT [44], TTSR [36], DCSR [31], MASA [21], C^2-matching [14], AMSA [34], DATSR [2], TDF [13] and LMR [40] as RefSR methods.

As reported in Table 1, CSSR-rec shows a better performance than the reconstruction loss only version of other methods. Especially, CSSR-rec gains over +0.3dB than other RefSR methods, on standard dataset CUFED5 in RefSR. Our CSSR also performs relatively better compared to other RefSR methods with full losses in terms of PSNR and SSIM. In particular, CSSR with full losses

Table 2. PSNR comparisons based on similarity level between LR and Ref images. Note that Ref images most similar to LR images are denoted by L1 and the least similar Ref images are denoted by L4. The best results are marked **in bold**.

Method	L1	L2	L3	L4	Average
	PSNR/SSIM	PSNR/SSIM	PSNR/SSIM	PSNR/SSIM	PSNR/SSIM
SRNTT-rec [44]	26.15/0.781	26.04/0.776	25.98/0.775	25.95/0.774	26.03/0.777
TTSR-rec [36]	26.99/0.800	26.74/0.791	26.64/0.788	26.58/0.787	26.74/0.792
DCSR-rec [31]	27.30/0.807	26.92/0.795	26.80/0.791	26.70/0.788	26.93/0.795
C^2-Matching-rec [14]	28.24/0.841	27.39/0.813	27.17/0.806	26.94/0.799	27.44/0.815
AMSA-rec [34]	28.58/0.849	27.52/0.816	27.25/0.809	27.04/0.803	27.60/0.819
DATSR-rec [2]	28.50/0.850	27.47/0.820	27.22/0.811	26.96/0.803	27.54/0.821
TDF-rec [13]	28.64/0.850	27.77/0.821	27.46/0.815	27.23/0.807	27.75/0.823
CSSR-rec	**29.03/0.858**	**27.98/0.828**	**27.71/0.820**	**27.45/0.811**	**28.04/0.829**

improves performance significantly on the CUFED5. Quantitative comparisons indicate that CSSR is generally superior to other methods.

Qualitative evaluations. We also provide visual results for the reconstruction loss only and full losses versions in Figure 5. It is confirmed that our method naturally transfers the details of the Ref image to the input LR image. In the left bottom of reconstruction loss version Fig. 5, CSSR delivers a similar clothing texture from the Ref, and produces a more detailed results image compared to other RefSR models. For another result in the right bottom of the full losses version, CSSR produces more accurate results exploiting letters in the Ref image.

4.3 Ablation study

Effects of similarity in reference images. We execute experiments to compare the performance according to the similarity level between LR and Ref images. We utilize CUFED5 dataset because it contains multiple Ref images with different similarity levels for a single LR image. In Table 2, our CSSR outperforms all RefSR methods even with different similarity levels. These results verify that CSSR is more robust even when Ref and LR images are less similar.

Effects of each module. In order to demonstrate the effectiveness of the proposed CSFT and HFFA modules, we conduct ablation studies by experiments by removing each module. As a baseline, we adopt a method that takes only warped Ref features using T obtained by (1) to transfer details of the Ref image. The performance of the baseline is reported in Table 3 (A), and compared with those of methods with the proposed HFFA and CSFT modules. As expected, our HFFA and CSFT modules increase the performance of the baseline RefSR model in terms of PSNR and SSIM. In addition, we qualitatively compare the results of variant versions of the proposed CSSR in Figure 6. We conduct those ablation studies on the CUFED5 dataset.

Cross-and self-feature transformer. As reported in Tab. 3 (D), the result with CSFT module has PSNR performance improvement of over +0.3 dB compared to (C). In Tab. 3 (B) shows that even in the absence of HFFA, the performance of the baseline with CSFT improves. Furthermore, we perform more

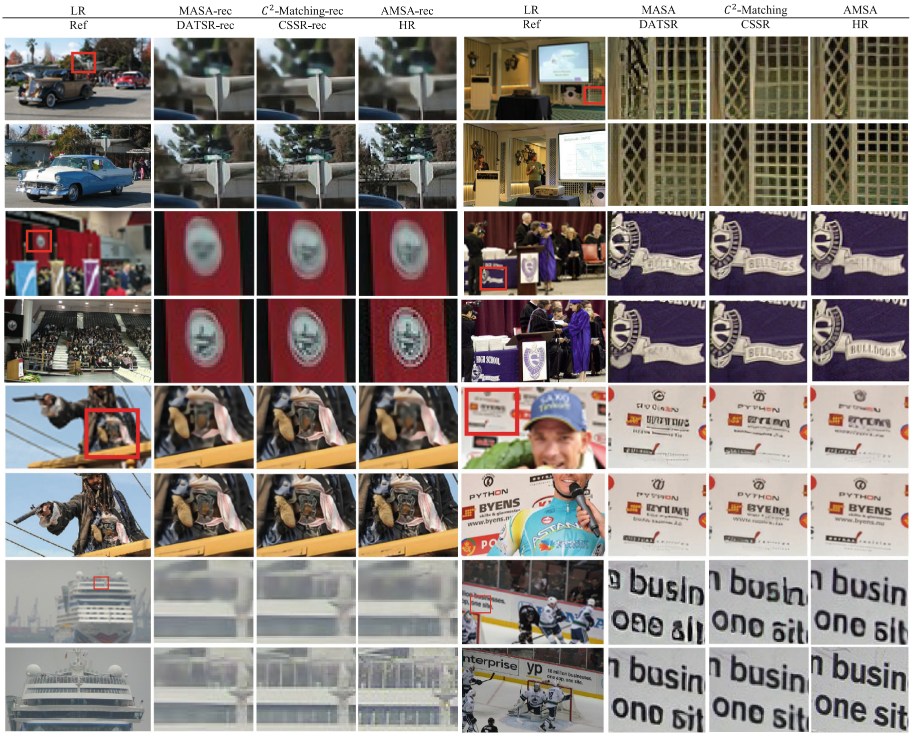


Fig. 5. Qualitative comparisons of RefSR models. The suffix '-rec' means reconstruction loss only version and without the suffix means full loss version.

in-depth experiments to check the effect of each submodule because the proposed CSFT module consists of the CFT and the SFT. For these experiments, we design several variations of CSFT as follows. On the top of the baseline model with HFFA, we gradually attach the CFT and the SFT to make variant versions. Since our CSFT of two CFTs, we name the first CFT as CFT1 and the second CFT as CFT2. As reported in Figure 7a, the performance increases as each module is added. The bright bar in Fig. 7a represents the results obtained using the entire CUFED5 dataset, while the dark bar corresponds to results using a subset of CUFED5 that includes complicated textures. A detailed description of this subset is provided in the supplementary material. Using the selected subset, we can further verify the effectiveness of our submodules. In Fig. 7b top, our model with the CFT1 produces detailed textures using Ref features. As can be seen in Fig. 7b middle, the model using all CFTs reconstructs more details due to the SISR features than the model with only the CFT1. Moreover, the effectiveness of the SFT module is validated the on Urban100 dataset that contains numerous patch recurrences. As shown in Fig. 7b bottom, CSSR with the SFT is more robust to repeating patterns than CSSR consisting of only two CFT modules. In Table 4, we explore the performance of the SFT and the CFT based on the similarity between Ref and LR images. The CFT1 improves performance because it utilizes rich texture from the Ref image in the case of L1. On the other hand, the CFT2 utilizes SISR features, resulting in a constant performance increase regardless of the similarity of the Ref image. Finally, the SFT performs better as it leverages internal features. These results illustrate that the CFT exploits features from both Ref images and restored images from the SISR model for aggregation, while the SFT leverages internal information by self-attention.

Table 3. Ablation study on each module. (A) The base. (B) With CSFT. (C) With HFFA. (D) Ours.

ID	HFFA	CSFT	PSNR/SSIM
(A)			28.53/0.844
(B)		✓	28.66/0.848
(C)	✓		28.72/0.849
(D)	✓	✓	**29.03/0.858**

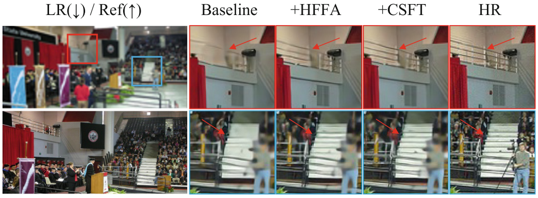
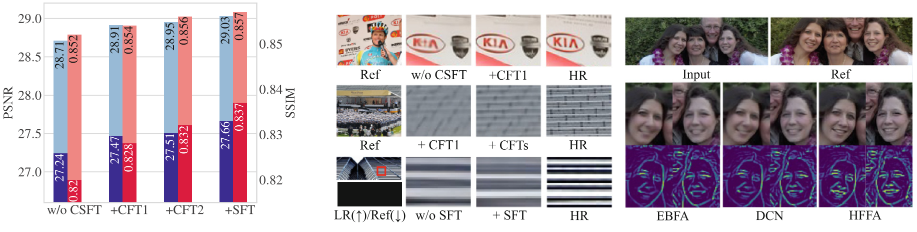


Fig. 6. Visualization of the effects of each module. The baseline with all modules shows the most detailed textures.

Table 4. Further analysis of CSFT submodules based on similarity level.

	L1	L2	L3	L4
	PSNR/SSIM	PSNR/SSIM	PSNR/SSIM	PSNR/SSIM
w/o CSFT	28.71/0.852	27.80/0.823	27.57/0.816	27.30/0.806
+CFT1	28.91/0.854	27.86/0.823	27.60/0.816	27.33/0.807
+CFT2	28.95/0.856	27.89/0.826	27.64/0.818	27.37/0.809
+SFT	**29.03/0.857**	**27.98/0.828**	**27.71/0.820**	**27.45/0.811**

High-frequency feature alignment module. In Tab. 3 (C), it is confirmed that adding HFFA improves the PSNR over the baseline performance. Also, as



(a) (b) (c)

Fig. 7. (a)Quantitative evaluation of CSFT. Submodules of CSFT are progressively attached to the baseline with HFFA module. Light color bar: entire CUFED5. Dark color bar: the subset of CUFED5. (b) Qualitative evaluation of CSFT. Effectiveness of the CFT1 (top). Effectiveness of the CFT1 and CFT2 (middle). Effectiveness of the SFT (bottom). (c) Qualitative evaluation of HFFA. The bottom results indicate the gradient images. Zoom in for the best quality.

Table 5. Ablation study on HFFA module. We replace HFFA module with the flow-based DCN [13] and EBFA [8], respectively.

	EBFA	DCN	HFFA
PSNR/SSIM	28.93/0.856	28.86/0.853	29.03/0.858
GMSD↓	0.1212	0.1221	0.1206

shown in Fig. 6, the high-frequency details are well restored with the HFFA than the baseline method. We perform more empirical studies to confirm the effectiveness of HFFA module. To verify that HFFA boosts high-frequency texture in the Ref image, we select gradient magnitude similarity deviation (GMSD) [35] as the gradient-based metric. We compare HFFA module to the existing flow-based DCN [2,13,14] and EBFA [8], which are recently adopted by many SR tasks. As reported in Table 5, CSSR with HFFA modules achieves superior performance than the other alignment modules. In Fig. 7c, the output with the gradient result of HFFA module contain more detailed texture.

5 Conclusion

In this paper, we have proposed a novel CSSR network for the RefSR task, which is robust even in cases where matching between input LR and Ref images is difficult. The cross-and self-feature transformer (CSFT) effectively aggregates features from both Ref and restored images as well as exploits internal information from the input LR image. In addition, we introduced a high-frequency feature alignment (HFFA) module to deliver the detailed residual features. The aforementioned three modules are seamlessly combined at multiple scales. Through various ablation studies, we verified the effectiveness of each proposed module in CSSR both quantitatively and qualitatively.

Acknowledgements. This work was supported by the Agency for Defense Development by the Korean Goverment(UI220050SD), National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIT) (No. 2022R1C1C1009334) and BK21 FOUR Program by Chungnam National University Research Grand, 2024.

References

1. Ba, J.L., Kiros, J.R., Hinton, G.E.: Layer normalization. arXiv preprint arXiv:1607.06450 (2016)
2. Cao, J., Liang, J., Zhang, K., Li, Y., Zhang, Y., Wang, W., Gool, L.V.: Reference-based image super-resolution with deformable attention transformer. In: Proceedings of the European Conference on Computer Vision. pp. 325–342 (2022)
3. Chang, H., Yeung, D.Y., Xiong, Y.: Super-resolution through neighbor embedding. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. I–I (2004)
4. Chen, H., Wang, Y., Guo, T., Xu, C., Deng, Y., Liu, Z., Ma, S., Xu, C., Xu, C., Gao, W.: Pre-trained image processing transformer. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 12299–12310 (2021)
5. Chen, X., Wang, X., Zhou, J., Qiao, Y., Dong, C.: Activating more pixels in image super-resolution transformer. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 22367–22377 (2023)
6. Dong, C., Loy, C.C., He, K., Tang, X.: Image super-resolution using deep convolutional networks. IEEE Trans. Pattern Anal. Mach. Intell. **38**(2), 295–307 (2015)
7. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. In: International Conference on Learning Representations (2021)
8. Dudhane, A., Zamir, S.W., Khan, S., Khan, F.S., Yang, M.H.: Burst image restoration and enhancement. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 5759–5768 (2022)
9. Freedman, G., Fattal, R.: Image and video upscaling from local self-examples. ACM TOG **30**(2), 1–11 (2011)
10. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. In: NeurIPS (2014)
11. Gulrajani, I., Ahmed, F., Arjovsky, M., Dumoulin, V., Courville, A.C.: Improved training of wasserstein gans. Proceedings of the Advances in Neural Information Processing Systems **30** (2017)
12. Huang, J.B., Singh, A., Ahuja, N.: Single image super-resolution from transformed self-exemplars. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 5197–5206 (2015)
13. Huang, Y., Zhang, X., Fu, Y., Chen, S., Zhang, Y., Wang, Y.F., He, D.: Task decoupled framework for reference-based super-resolution. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 5931–5940 (2022)
14. Jiang, Y., Chan, K.C., Wang, X., Loy, C.C., Liu, Z.: Robust reference-based super-resolution via c2-matching. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 2103–2112 (2021)
15. Kim, K.I., Kwon, Y.: Single-image super-resolution using sparse regression and natural image prior. IEEE Trans. Pattern Anal. Mach. Intell. **32**(6), 1127–1133 (2010)

16. Ledig, C., Theis, L., Huszár, F., Caballero, J., Cunningham, A., Acosta, A., Aitken, A., Tejani, A., Totz, J., Wang, Z., et al.: Photo-realistic single image super-resolution using a generative adversarial network. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 4681–4690 (2017)
17. Liang, J., Cao, J., Sun, G., Zhang, K., Van Gool, L., Timofte, R.: Swinir: Image restoration using swin transformer. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 1833–1844 (2021)
18. Liang, J., Sun, G., Zhang, K., Van Gool, L., Timofte, R.: Mutual affine network for spatially variant kernel estimation in blind image super-resolution. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 4096–4105 (2021)
19. Lim, B., Son, S., Kim, H., Nah, S., Mu Lee, K.: Enhanced deep residual networks for single image super-resolution. In: Proceedings of the IEEE/CVF International Conference on Computer Vision Workshop. pp. 136–144 (2017)
20. Liu, J., Zhang, W., Tang, Y., Tang, J., Wu, G.: Residual feature aggregation network for image super-resolution. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 2359–2368 (2020)
21. Lu, L., Li, W., Tao, X., Lu, J., Jia, J.: Masa-sr: Matching acceleration and spatial adaptation for reference-based image super-resolution. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 6368–6377 (2021)
22. Luo, Z., Huang, H., Yu, L., Li, Y., Fan, H., Liu, S.: Deep constrained least squares for blind image super-resolution. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 17642–17652 (2022)
23. Matsui, Y., Ito, K., Aramaki, Y., Fujimoto, A., Ogawa, T., Yamasaki, T., Aizawa, K.: Sketch-based manga retrieval using manga109 dataset. Multimedia Tools and Applications **76**(20), 21811–21838 (2017)
24. Mei, Y., Fan, Y., Zhou, Y.: Image super-resolution with non-local sparse attention. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 3517–3526 (2021)
25. Shim, G., Park, J., Kweon, I.S.: Robust reference-based super-resolution with similarity-aware deformable convolution. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 8425–8434 (2020)
26. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)
27. Sun, J., Xu, Z., Shum, H.Y.: Image super-resolution using gradient profile prior. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 1–8 (2008)
28. Sun, L., Hays, J.: Super-resolution from internet-scale scene matching. In: IEEE International Conference on Computational Photography. pp. 1–12 (2012)
29. Tai, Y., Yang, J., Liu, X.: Image super-resolution via deep recursive residual network. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 3147–3155 (2017)
30. Timofte, R., De Smet, V., Van Gool, L.: Anchored neighborhood regression for fast example-based super-resolution. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 1920–1927 (2013)
31. Wang, T., Xie, J., Sun, W., Yan, Q., Chen, Q.: Dual-camera super-resolution with aligned attention modules. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 2001–2010 (2021)
32. Wang, X., Yu, K., Wu, S., Gu, J., Liu, Y., Dong, C., Qiao, Y., Change Loy, C.: Esrgan: Enhanced super-resolution generative adversarial networks. In: Proceedings of the European Conference on Computer Vision Workshop. pp. 0–0 (2018)

33. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: from error visibility to structural similarity. IEEE TIP **13**(4), 600–612 (2004)
34. Xia, B., Tian, Y., Hang, Y., Yang, W., Liao, Q., Zhou, J.: Coarse-to-fine embedded patchmatch and multi-scale dynamic aggregation for reference-based super-resolution. In: AAAI (2022)
35. Xue, W., Zhang, L., Mou, X., Bovik, A.C.: Gradient magnitude similarity deviation: A highly efficient perceptual image quality index. IEEE TIP **23**(2), 684–695 (2013)
36. Yang, F., Yang, H., Fu, J., Lu, H., Guo, B.: Learning texture transformer network for image super-resolution. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 5791–5800 (2020)
37. Yang, J., Lin, Z., Cohen, S.: Fast image super-resolution based on in-place example regression. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 1059–1066 (2013)
38. Yang, J., Wright, J., Huang, T.S., Ma, Y.: Image super-resolution via sparse representation. IEEE TIP **19**(11), 2861–2873 (2010)
39. Zamir, S.W., Arora, A., Khan, S., Hayat, M., Khan, F.S., Yang, M.H.: Restormer: Efficient transformer for high-resolution image restoration. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 5728–5739 (2022)
40. Zhang, L., Li, X., He, D., Li, F., Ding, E., Zhang, Z.: Lmr: A large-scale multi-reference dataset for reference-based super-resolution. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 13118–13127 (2023)
41. Zhang, L., Li, X., He, D., Li, F., Wang, Y., Zhang, Z.: Rrsr: Reciprocal reference-based image super-resolution with progressive feature alignment and selection. In: Proceedings of the European Conference on Computer Vision. pp. 648–664 (2022)
42. Zhang, X., Zeng, H., Guo, S., Zhang, L.: Efficient long-range attention network for image super-resolution. In: Proceedings of the European Conference on Computer Vision. pp. 649–667 (2022)
43. Zhang, Y., Li, K., Li, K., Wang, L., Zhong, B., Fu, Y.: Image super-resolution using very deep residual channel attention networks. In: Proceedings of the European Conference on Computer Vision. pp. 286–301 (2018)
44. Zhang, Z., Wang, Z., Lin, Z., Qi, H.: Image super-resolution by neural texture transfer. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 7982–7991 (2019)
45. Zheng, H., Ji, M., Wang, H., Liu, Y., Fang, L.: Crossnet: An end-to-end reference-based super resolution network using cross-scale warping. In: Proceedings of the European Conference on Computer Vision. pp. 88–104 (2018)
46. Zhou, Y., Li, Z., Guo, C.L., Bai, S., Cheng, M.M., Hou, Q.: Srformer: Permuted self-attention for single image super-resolution. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 12780–12791 (2023)

Transformer-Based Depth Optimization Network for RGB-D Salient Object Detection

Lu Li[1], Yanjiao Shi[1](✉), Jinyu Yang[1], Qiangqiang Zhou[2], Qing Zhang[1], and Liu Cui[1]

[1] Shanghai Institute of Technology, Shanghai, China
shiyanjiao616@163.com
[2] Jiangxi Normal University, Nanchang, China

Abstract. Given the increasing emphasis on multimodal data analytics, depth maps have been employed for Salient Object Detection (SOD) task. RGB-D SOD task utilizes the spatial structure information in the depth maps to improve detection accuracy. In this paper, we propose a Transformer-based Depth Optimization Network (DONet) for RGB-D SOD task. A depth feature optimization and integration module (DOIM) is first designed to maximize the auxiliary effect of depth information. In DOIM, high-quality depth information is retained and low-quality information is discarded conversely. Then aiming to comprehensive detail complement, the context supplement modules (CSMs) are configured to absorb features of adjacent layers to refine the features adequately. In addition, for global information exploration, we deploy a location perception guider (LPG) to guide our model to explore the location of salient objects accurately. Based on the wide application of Transformer, the Pyramid Vision Transformer with less computational requirement and equal efficiency is chosen to balance performance and computational cost. Experiments on five widely-used datasets show that the proposed network achieves significant advantage compared to 13 state-of-the-art methods.

Keywords: Image processing · RGB-D salient object detection · Deep learning · Depth feature optimization · Transformer

1 Introduction

Salient object detection (SOD) is a crucial computer vision task aimed to identify and highlight the most prominent objects within an image [13]. Traditional methods for SOD typically relied on factors such as image features, color, and texture. However, relying solely on RGB information is not enough to resist the interference of some complex factors such as illumination changes and shadows. This limitation can be compensated by depth maps that provide the distance of the object to the lens. With the development of RGB-D sensor technology, depth maps can be easily acquired from devices such as stereo cameras and smart phones, which promoted the wide application of RGB-D data in computer vision. Concurrently, within the prevailing trend of multimodal data analysis, an increasing number of researchers have been dedicated to RGB-D SOD task, which leverages the spatial structural information of depth maps to assist model in performing detection task.

A. Antonacopoulos et al. (Eds.): ICPR 2024, LNCS 15321, pp. 435–450, 2025.
https://doi.org/10.1007/978-3-031-78305-0_28

The advent of deep learning technology [11] has led to significant advancements in salient object detection. Most of the prior RGB-D SOD methods based on convolution neural networks (CNNs) have achieved superior performance. However, CNNs have relatively weaker capture of global relationships within the overall scene. In contrast, Transformer [25] is adept at handling positional information within sequences, which is crucial for a nuanced understanding of object context in salient object detection. In the past two years, Transformer has been gradually applied to RGB-D SOD task. Most Transformer-based models show significant advantages, which motivates us to adopt Transformer as the backbone network.

In RGB-D SOD task, it is necessary to focus on the fusion of depth maps and RGB images. However, most of the previous works ignored the impact of different qualities of depth information. The quality level of depth information is uneven, high-quality depth information should be retained while low-quality depth information should be discarded. Based on this, we design a depth feature optimization and integration module (DOIM) to further optimize and fuse depth and RGB information. Specifically, we first enhance the preliminary extracted depth features, then adopt the attention mechanisms to filter them before merging with RGB features, so as to maximize the auxiliary effect of the depth information. Detailed information contains crucial content. Whereas, after the depth feature optimization, many fine-grained details are inevitably lost. Three context supplement modules (CSMs) are configured in the proposed network to enable detailed information exploration. The overall strategy of CSM is absorbing the features of previous layers and subsequent layers (for the first layer, only the subsequent layer is introduced) to supplement the details of the current layer. Furthermore, we improve the effect of detail complement by the combination of spatial attention and channel attention mechanisms. In SOD task, accurate positioning plays a vital role in detecting salient objects. Generally speaking, high-level features contain rich global information, which contributes to identifying the locations of salient objects. Based on this, a location perception guider (LPG) is designed on the features of the last two layers. Different from previous works, we leverage the self-attention mechanism to model the global relationships in high-level features. To guide the proposed network to locate salient objects more accurately and adequately, the features with location-aware information are sent vertically to the CSM and horizontally to the fusion block.

Our main contributions can be summarized as follows:

- We propose a Transformer-based depth optimization network (DONet) for RGB-D SOD task, which focuses on the optimization of depth information, detail supplement of features and perception of salient objects' location.
- To maximize the auxiliary effect of depth information, we design a depth feature optimization and integration module (DOIM). The module enhances and filters depth information, then aggregates it with the RGB information.
- For comprehensive detail complement, context supplement modules (CSMs) are configured to absorb features of adjacent layers, expanding the coverage of interactions to refine the features adequately.
- We deploy a location perception guider (LPG) with self-attention mechanism for global information exploration of high-level features in order to accurately locate salient objects.

2 Related Works

RGB-D salient object detection task utilizes the combination of RGB image and depth map to identify salient object. RGB images contain appearance and texture information, while depth maps provide spatial structural information. The fusion of RGB and depth features has been a crucial issue in RGB-D SOD. According to the order of fusion operations in the whole decoding process, some RGB-D SOD methods employed early fusion [4,38], while others utilized dual-branch networks for intermediate fusion [1,15] and late fusion [6,21]. To address the problem of low-quality depth maps, D3Net [9] employed a gating mechanism to filter out inferior depth maps, EF-Net [3] enhanced depth maps using RGB hint maps. Considering the importance of both enhancement and filtering of depth features, in this paper, we adopt the strategy of enhancing first and filtering subsequently to maximize the auxiliary effect of depth information.

Transformer has achieved significant success in natural language processing (NLP) and subsequently been widely applied in computer vision tasks. An increasing number of researchers have adopted Transformer in RGB-D SOD task. Liu et al. [20] inserted a triplet transformer module into the CNNs-based backbone network to enhance high-level features. Subsequently, Cong et al. [8] reversed the primary and secondary relationship of the two architectures, using CNNs-based model as auxiliary component of Swin Transformer-based architecture to optimize global context and local details. Liu et al. [19] proposed to use a cross-modal RGB-D and RGB-T SOD fusion network based on Swin Transformer. The model proposed by Wang et al. [28] was also based on Swin Transformer but with additional edge guidance. In prior RGB-D SOD methods based on Transformer, Swin Transformer [18] was predominantly utilized. However, the substantial computational demand of Swin Transformer often results in prolonged training duration. Pyramid Vision Transformer (PVT) [30], another Transformer commonly used in computer vision, disposes features at different scales through a pyramid structure to better capture the information. Compared with Swin Transformer, PVT has a small computational demand but equal efficiency. Therefore, considering the balance between performance and computational cost, we chose PVT as the backbone network.

3 Methodology

3.1 Network Overview

The overall framework of the proposed DONet is shown in Figure 1. According to encoder-decoder strategy, DONet consists of backbone network, depth feature optimization and integration module (DOIM), context supplement module (CSM), location perception guider (LPG) and fusion block.

To fit the configuration of the backbone network, both the RGB and depth images are resized to 224×224, then fed into the shared Pyramid Vision Transformer for feature extraction to get feature RGB_i and D_i of the i^{th} layer, $i \in \{1, 2, 3, 4\}$. The DOIM allows further feature extraction and selecting high-quality depth information to achieve maximum auxiliary effect. The CSM introduces the information of neighboring layers to optimize features, while LPG strengthens the learning of high-level features to capture global information so as to locate salient objects accurately. After fusion and linear layer, four refined maps are obtained, marked as Out_i, $i \in \{1, 2, 3, 4\}$, which are supervised by the ground truth maps. The Out_1 is picked to be the final prediction map in the inference stage.

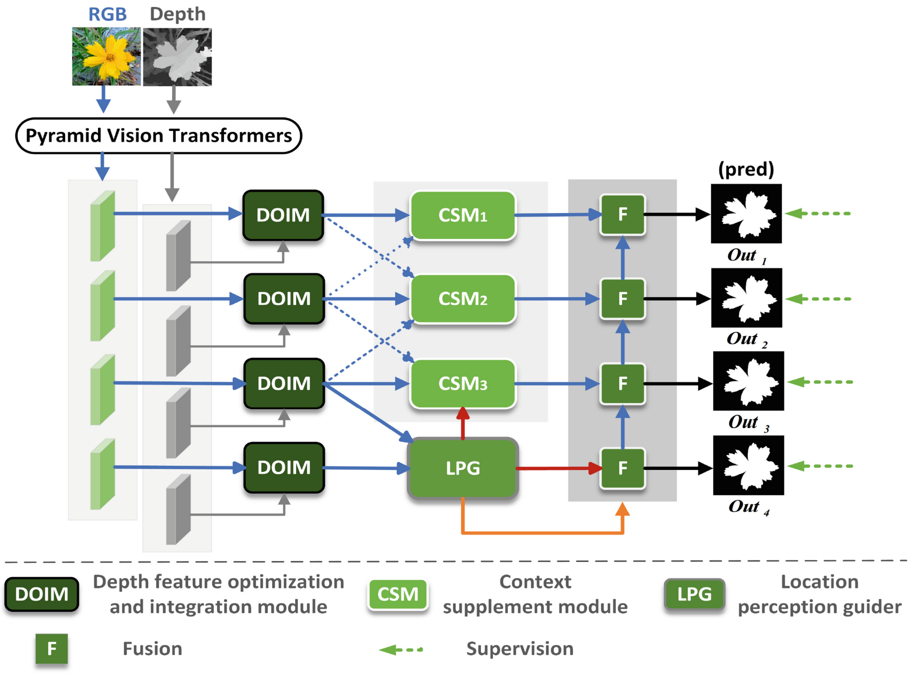

RGB
Depth
Pyramid Vision Transformers
DOIM
DOIM
DOIM
DOIM
CSM1
CSM2
CSM3
LPG
F
F
F
F
(pred)
Out 1
Out 2
Out 3
Out 4
DOIM
Depth feature optimization and integration module
CSM
Context supplement module
LPG
Location perception guider
F
Fusion
Supervision

Fig. 1. The overall framework of the proposed **DONet**. The input images are resized to 224×224. DONet consists of backbone network (PVT), depth feature optimization and integration module (DOIM), context supplement module (CSM), location perception guider (LPG) and fusion block. We take the Out_1 as the final prediction map.

3.2 Depth Feature Optimization and Integration Module

After feature extraction through backbone network, we need pay more attention to achieve efficient interaction between multi-level features of RGB and depth modality in the encoding stage. Both the RGB and depth features extracted from the backbone network are pretty coarse, so direct fusion will lead to inefficient decoding after grafting. The grafting efficiency can be significantly improved by filtering out low-quality depth features. In view of this, we propose a depth feature optimization and integration module (DOIM). In DOIM, dual flow information is enhanced through asymmetric convolutions stacked in parallel, while depth information is refined by attention mechanisms.

As shown in Figure 2, our proposed DOIM contains multi-scale features enhancement (MSE) and depth optimization. For MSE, we first utilize a 1 × 1 convolution layer on the input feature F_i of the i^{th} layer to obtain the F_i^d with 64 channels. Then the MSE is equipped with four parallel convolution groups, each of which contains connected 1×3 and 3×1 asymmetric convolutions. The output from each group is passed to the subsequent group for element-wise addition. We define the output of each group as M_i^k, $k \in \{1, 2, 3, 4\}$, which is formulized as:

$$M_i^k = \begin{cases} CBR_{3\times1}\left(CBR_{1\times3}\left(F_i^d\right)\right) & k = 1 \\ CBR_{3\times1}\left(CBR_{1\times3}\left(F_i^d + M_i^{k-1}\right)\right) & k = 2, 3, 4 \end{cases} \tag{1}$$

where $CBR_{3\times1}(\cdot)$ and $CBR_{1\times3}(\cdot)$ indicate 3 × 1 and 1 × 3 convolution followed by batch normalization and relu layer, respectively. After that, we concatenate the outputs of four groups followed by a 3 × 3 convolution and make a residual addition with the F_i^d. The output of each MSE is defined as MSE_i, which is formulized as:

$$MSE_i = CBR_{3\times3}\left(Cat\left(M_i^1, M_i^2, M_i^3, M_i^4\right)\right) + F_i^d, \tag{2}$$

where $CBR_{3\times3}(\cdot)$ indicates 3 × 3 convolution followed by batch normalization and relu layer, $Cat(\cdot)$ indicates channel-wise concatenation. For the input RGB feature RGB_i and depth feature D_i, we get the enhanced feature MSE_i^{RGB} and MSE_i^D, respectively. Before the aggregation, depth features are filtered. Specifically, for MSE_i^D, we use channel attention to obtain the corresponding attention map MSE_i^{CA}, followed by spatial attention to obtain $MSE_i^{D-attention}$, which is formulized as:

$$MSE_i^{CA} = MSE_i^D * CA\left(MSE_i^D\right), \tag{3}$$

$$MSE_i^{D-attention} = MSE_i^{CA} * SA\left(MSE_i^{CA}\right), \tag{4}$$

where $*$ denotes element-wise multiplication, $CA(\cdot)$ and $SA(\cdot)$ denote channel attention and spatial attention, respectively. At last, we aggregate the depth and RGB features, which is formulated as:

$$DOIM_i = MSE_i^{RGB} + MSE_i^{D-attention}, \tag{5}$$

where $DOIM_i$ is the output of the whole depth feature optimization and integration module of the i^{th} layer. The aggregated features will be fed into the subsequent decoding modules CSM and LPG.

3.3 Context Supplement Modules

The proposed context supplement modules (CSMs) connects DOIM and LPG, with the purpose of refining the features of the previous stage and absorbing the location-aware information. After the depth information decoding and aggregation in the previous stage, certain detailed information are unavoidably lost. In order to comprehensively supplement detailed information, we configures three CSMs, namely CSM_1, CSM_2, and CSM_3. As illustrated in Figure 1, each CSM assimilates features from the current layer and its adjacent layers. This cross-linking mechanism expands the coverage

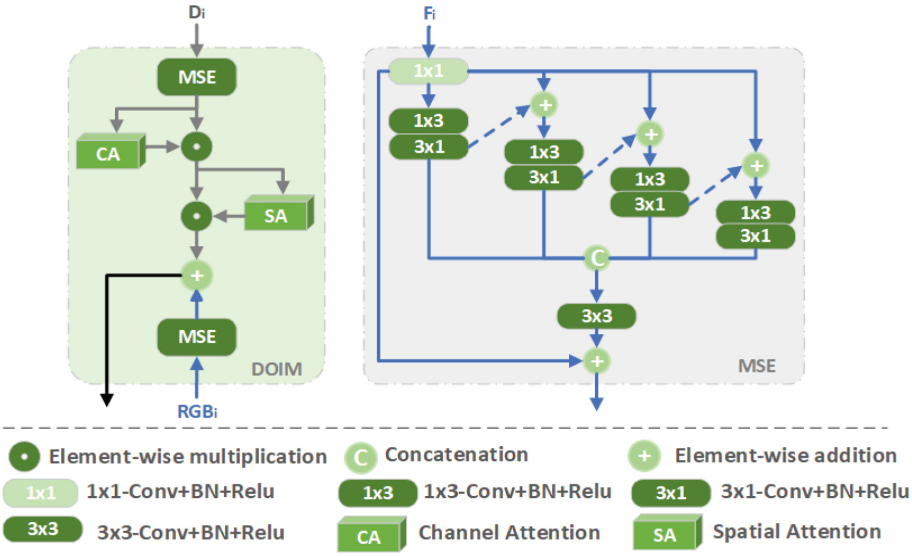


Fig. 2. The structure of the proposed depth feature optimization and integration module (DOIM).

of feature interaction, establishing a close relationship between features at different layers. Employing the strategy analogous to DOIM, we incorporates attention mechanisms, which allows self-adaptive selection of more valuable content from features of adjacent layers. Notably, CSM_3 further integrates location-aware information from the LPG by substituting the fourth-layer feature with the output of LPG. We define the processing of CSM as $F(\cdot)$, which is formulated as follows:

$$CSM_i = \begin{cases} F\left(DOIM_1, DOIM_2\right) & i = 1 \\ F\left(DOIM_1, DOIM_2, DOIM_3\right) & i = 2 \\ F\left(DOIM_2, DOIM_3, LPG_3\right) & i = 3 \end{cases} \tag{6}$$

As shown in Figure 3 , the features of previous and subsequent layers are introduced in addition to current layer for CSM_2 and CSM_3, while only the subsequent layer is introduced for CSM_1. We extract the previous attention map P_{mi}, subsequent attention map S_{mi} and current enhancement map C_{mi} by the combined operations of subtle channel attention (CA) , spatial attention (SA) and convolutions, which can be formulized as:

$$P_{mi} = Conv_{3\times3}\left(DOIM_{i-1}\right) * SA\left(Conv_{3\times3}\left(DOIM_{i-1}\right)\right), \qquad i = 2, 3 \tag{7}$$

$$S_{mi} = \begin{cases} Conv_{3\times3}\left(DOIM_{i+1}\right) * CA\left(Conv_{3\times3}\left(DOIM_{i+1}\right)\right) & i = 1, 2 \\ Conv_{3\times3}\left(LPG_3\right) * CA\left(Conv_{3\times3}\left(LPG_3\right)\right) & i = 3 \end{cases} \tag{8}$$

$$C_{mi} = Conv_{3\times3}\left(DOIM_i\right), \qquad i = 1, 2, 3 \tag{9}$$

where $*$ denotes element-wise multiplication, $Conv_{3\times3}(\cdot)$ denotes 3×3 convolution. After concatenating P_{mi}, S_{mi} and C_{mi}, we obtain map D_{mi} by channel adjustment and sent it to convolution layer. At last, the final output of CSM is generated by residual addition with $DOIM_i$ and convolution layer.

$$D_{mi} = \begin{cases} Conv_{1\times1}\left(Cat\left(S_{mi}, C_{mi}\right)\right), & i = 1 \\ Conv_{1\times1}\left(Cat\left(P_{mi}, S_{mi}, C_{mi}\right)\right), & i = 2, 3 \end{cases} \tag{10}$$

$$CSM_i = CBR_{3\times3}\left(CBR_{3\times3}\left(D_{mi}\right) + DOIM_i\right), \qquad i = 1, 2, 3 \tag{11}$$

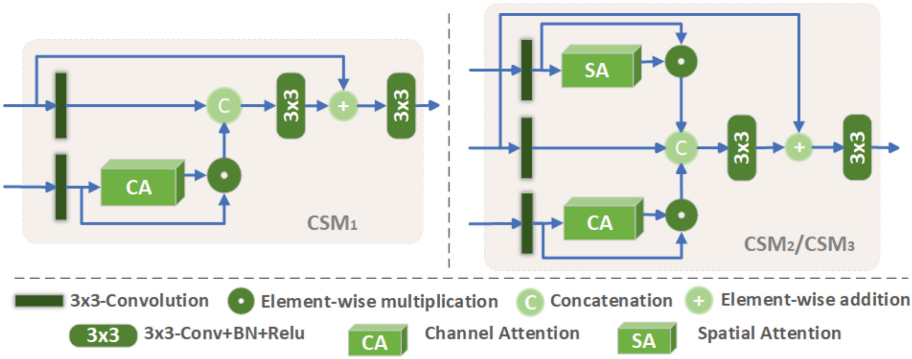


Fig. 3. The structure of proposed context supplement modules (CSMs).

3.4 Location Perception Guider

The mining of global information has always been a key issue in SOD task. Unlike some existing SOD methods applying semantic segmentation modules such as ASPP [2] or PSP modules [37] directly, we design a self-attention mechanism based location perception guider (LPG). Considering that there is rich global information in the high-level features, we choose the last two layers as the benchmark for object positioning. Details of LPG are illustrated in Figure 4. First, we resize $DOIM_4$ to align it with $DOIM_3$, then flatten the $DOIM_3 \in v^{C\times H\times W}$ to $DOIM_3^f \in v^{1\times C\times HW}$. Perform the same operations on $DOIM_4$ as described above, then we transpose it to get $DOIM_4^{ft} \in v^{1\times HW\times C}$. A matrix multiplication is performed on the linearized $DOIM_4^{ft}$and $DOIM_3^f$ to obtain Y, which can be formulized as:

$$Y = L\left(DOIM_4^{ft}\right) \times DOIM_3^f, \tag{12}$$

where × indicates matrix multiplication, $L(\cdot)$ indicates the linear projection. We perform a softmax operation on Y, followed by matrix multiplication with $DOIM_3^f$ and $DOIM_4^f$, respectively. Then we reshape them back to size of $v^{C\times H\times W}$to obtain S_3 and S_4 , which can be formulized as:

$$S_3 = reshape\left(DOIM_3^f \times softmax\left(Y\right)\right), \tag{13}$$

$$S_4 = reshape\left(DOIM_4^f \times softmax\left(Y^t\right)\right), \tag{14}$$

where Y^t indicates the the transposed Y. On the basis of S_3 and S_4, the channel attention mechanism is adopted, which allows our model to explore the attention weight from the channel dimension and adjust the features adaptively. We multiply S_3 by itself after sending it into the channel attention component, and then make addition supplement with $DOIM_3$ before feeding the convolutional block. The same operation is applied to S_4. At last, we input them into the convolutional layer with double residual connections to get the output of LPG. The process can be expressed as:

$$C_3 = CBR2_{3\times 3}\left(CA\left(S_3\right) * S_3 + DOIM_3\right), \tag{15}$$

$$LPG_3 = CBR2_{3\times 3}\left(CBR2_{3\times 3}\left(C_3 + DOIM_3\right) + C_3\right), \tag{16}$$

$$C_4 = CBR2_{3\times 3}\left(CA\left(S_4\right) * S_4 + reshape\left(DOIM_4\right)\right), \tag{17}$$

$$LPG_4 = CBR2_{3\times 3}\left(CBR2_{3\times 3}\left(C_4 + reshape(DOIM_4)\right) + C_4\right), \tag{18}$$

where $CBR2_{3\times 3}(\cdot)$ indicates 2 stacked 3×3 convolution followed by batch normalization and relu layer.

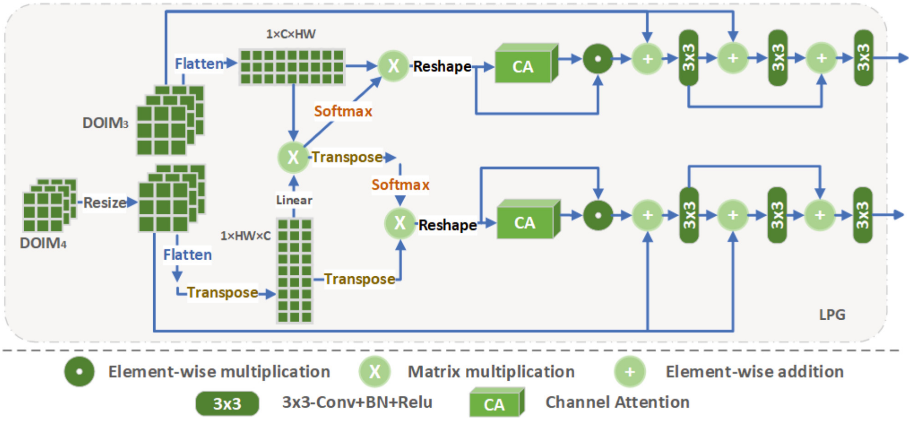


Fig. 4. Details of the proposed location perception guider (LPG).

3.5 Loss Function

To improve the accuracy of location perception guider, we apply additional supervision to the boundaries. Referring to the work [35], the boundary DICE (BD) loss is employed in our training process. The overall strategy of BD loss is to compute the DICE loss [16] value on the distensible boundaries. For ground truth maps and prediction maps, the corresponding thin boundary maps are obtained by dilation and erosion operations, then max-pooling is performed to enlarge the boundary region. After that, the DICE loss is calculated on this thick boundary. L_{bd} denotes the BD loss, which can be formulated as:

$$L_{bd}\left(P^b, G^b\right) = 1 - \frac{1 + \sum_{i=1,j=1}^{H,W} 2 \times P_{ij}^b \times G_{ij}^b}{1 + \sum_{i=1,j=1}^{H,W} P_{ij}^b + G_{ij}^b}, \tag{19}$$

$$P_{ij}^b = max\left(P_{A_{ij}}^{b,thin}\right), \tag{20}$$

$$G_{ij}^b = max\left(G_{A_{ij}}^{b,thin}\right), \tag{21}$$

where $max(\cdot)$ means the max-pooling operation, $G_{A_{ij}}^{b,thin}$and $P_{A_{ij}}^{b,thin}$ represent the thin boundary map of the ground truth map G and the prediction map P, respectively, A_{ij} represents the pooling area that surrounds the pixel (i, j). In addition, we also apply pixel-level weighted BCE loss [31] to pay more attention to hard pixels, and map-level IOU loss [33] to focus on the global structure of the map. Our proposed network includes four supervised outputs in total–the final prediction map (Out_1) and three refined maps (Out_2, Out_3 and Out_4). We combine the three kinds of loss functions mentioned above as the total loss for the four supervisions, which can be formulated as:

$$L_i = L_{wbce}\left(Out_i, G\right) + L_{IOU}\left(Out_i, G\right) + L_{bd}\left(Out_i^b, G^b\right) \quad i \in \{1, 2, 3, 4\}. \tag{22}$$

According to different levels of refinement, we assign decreasing attention to the four supervisions. The sum of losses over the whole training process can be expressed as:

$$L_{Total} = \sum_{i=1}^{4} \frac{1}{2^{i-1}} L_i. \tag{23}$$

4 Experiment

4.1 Datasets and Evaluation Metrics

We evaluate the proposed network on 5 public RGB-D SOD datasets. NLPR [23] has 1,000 images with single or multiple salient objects. NJU2K [14] includes 2,003 stereo image pairs and ground-truth maps in complex scenes. STERE [22] incorporates 1,000 pairs of binocular images from the Internet. SIP [9] contains 1,000 high-resolution images of multiple salient persons. SSD [40] holds 80 images from stereo movies with various scenes. We use the same training dataset as in [9] and [5], which consists of 1,485 samples from the NJU2K dataset and 700 samples from the NLPR dataset. The training dataset is augmented by random cropping, flipping and rotation operations to prevent the phenomenon of overfitting.

The performance of our model and other methods are evaluated through five widely recognized metrics including S-measure (*Sm*) [29], mean F-measure (*Fm*) [29], weighted F-measure (*Wgt-F*) [29], E-measure (*Em*) [29] and mean absolute error (*MAE*) [29]. To ensure a fair comparison, evaluation is performed using saliency maps provided by the authors or acquired by testing with the official codes.

4.2 Implementation Details

The proposed model is deployed in the Pytorch framework and trained on an Nvidia RTX 3080Ti GPU. The backbone PVT is initialized with parameters pretrained by ImageNet. We employ the Adam optimizer to train our model with a batch size of 10 and an initial learning rate 1e-4. The training converges within 68 epochs.

4.3 Compared with the State-of-the-art Methods

We conduct a comparison with 13 state-of-the-art models, including BBSNet [34], DSA^{2}F [24], TriTransNet [20], VST [17], DCMF [26], JL-DCF [10], DCFNet [12], C^{2}DFNet [36], MVSalNet [39], CIRNet [7], PICRNet [8], HIDANet [32], and AMINet [27]. Among them, VST is a transformer-based model, and TriTransNet and PICRNet are models with Transformer-assisted CNNs and CNNs-assisted Transformer, respectively.

1. **Quantitative evaluation:** The quantitative results of the proposed DONet on five commonly used datasets are presented in table 1. Clearly, our model achieves a significant advantage, as evidenced by the fact that almost all the values of our model are either optimal or suboptimal. Specifically, on the NLPR and NJU2K datasets, our DONet achieves optimality on all five metrics. Compared with the 2023 PICRNet [8], our model improves the *MAE* of these three datasets by 13.8%, 15.8% and 6.5%, respectively.
2. **Qualitative evaluation:** Figure 5 presents some examples of visual comparisons with other methods in several challenging scenarios. It can be observed that for objects with similar colors in edge and background (*i.e.*, b and e), the location is annotated more comprehensively in our prediction maps. On irregular and densely curved objects (*i.e.*, a, c, d and f), our prediction exhibit higher boundary precision and more detailed descriptions. Even in cases where the quality of the depth map is poor (*e.g.*, d and e), the proposed model accurately reconstructs the primary content and boundary details of the salient object.

4.4 Ablation Study

We conduct ablation studies on NJU2K and NLPR test datasets to verify the contribution of each proposed module or component in the proposed DONet.

Effectiveness of General Structure We verify the effectiveness of DOIM, CSMs, and LPG in Table 2. ID 0: DONet with full structure; ID 1: DOIM replaced by element-wise addition; ID 2 and ID 3: CSMs and LPG removal, respectively.

Compared with ID 0, the deteriorating data in ID 1 shows the better performance of DOIM, which verifies that the optimization of depth features before the aggregation of two-stream features is necessary for subsequent decoding. The strategy of introducing the features of adjacent layers is commonly used in RGB SOD task. Based on this, we design CSMs and apply it to RGB-D SOD task. Observably, the improvement in

Table 1. Quantitative comparisons with 13 state-of-the-art methods. The optimal and suboptimal values are highlighted in red and blue, respectively. "↑" means that higher value is better. "↓" means that lower value is better.

Methods	Year/Pub	NJU2K					NLPR					SIP					STERE					SSD				
		$MAE\downarrow$	$Fm\uparrow$	$Em\uparrow$	$Sm\uparrow$	$Wgt\text{-}F\uparrow$	$MAE\downarrow$	$Fm\uparrow$	$Em\uparrow$	$Sm\uparrow$	$Wgt\text{-}F\uparrow$	$MAE\downarrow$	$Fm\uparrow$	$Em\uparrow$	$Sm\uparrow$	$Wgt\text{-}F\uparrow$	$MAE\downarrow$	$Fm\uparrow$	$Em\uparrow$	$Sm\uparrow$	$Wgt\text{-}F\uparrow$	$MAE\downarrow$	$Fm\uparrow$	$Em\uparrow$	$Sm\uparrow$	$Wgt\text{-}F\uparrow$
BBSNet	21 TIP	.035	.912	.942	.921	.884	.023	.904	.953	.93	.88	.055	.885	.917	.879	.83	.041	.896	.941	.908	.858	.044	.869	.92	.882	.811
DSA^{2}F	21 CVPR	.04	.907	.936	.904	.882	.024	.904	.951	.919	.88	-	-	-	-	-	-	-	-	-	-	.048	.864	.911	.877	.824
TriTrans	21 ACMMM	.03	.924	.954	.919	.906	.02	.916	.964	.928	.902	.043	.905	.929	.866	.864	.033	.899	.95	.908	.882	.041	.876	.935	.881	.842
VST	21 ICCV	.034	.908	.943	.922	.888	.023	.907	.956	.931	.887	.04	.906	.941	.903	.872	.038	.889	.942	.913	.866	.044	.863	.922	.889	.829
JL-DCF	21 TPAMI	.045	.879	.924	.893	.853	.022	.900	.955	.925	.882	.049	.886	.923	.88	.844	.04	.88	.937	.903	.857	.053	.831	.899	.860	.782
DCMF	22 TIP	.045	.885	.919	.909	.856	.029	.883	.94	.922	.856	-	-	-	-	-	.043	.881	.93	.91	.849	.053	.852	.895	.883	.803
DCFNet	22 CVPR	.039	.905	.94	.903	.876	.024	.903	.955	.922	.884	.052	.888	.92	.873	.84	.037	.898	.943	.906	.872	.054	.846	.905	.852	.789
C^{2}DFNet	22 TMM	.039	.906	.94	.908	.878	.022	.912	.96	.928	.888	.053	.88	.917	.872	.83	.038	.892	.941	.902	.862	.048	.866	.918	.872	.816
MVSalNet	22 ECCV	.036	.911	.937	.911	.884	.021	.912	.957	.93	.893	-	-	-	-	-	.035	.904	.945	.913	.878	-	-	-	-	-
CIRNet	22 TIP	.035	.916	.943	.925	.891	.023	.908	.955	.933	.884	.052	.89	.918	.888	.84	-	-	-	-	-	-	-	-	-	-
PICRNet	23 ACMMM	.029	.926	.951	.927	.909	.019	.922	.967	.935	.907	-	-	-	-	-	.031	.912	.952	.921	.892	-	-	-	-	-
HIDANet	23 TIP	.029	.928	.953	.926	.910	.021	.916	.961	.930	.901	.044	.907	.927	.892	.866	.035	.904	.945	.911	.881	.047	.859	.917	.87	.819
AMINet	23 ACMTM	.036	.916	.946	.906	.889	.025	.9	.955	.908	.872	-	-	-	-	-	.04	.892	.939	.890	.863	-	-	-	-	-
Ours	-	.025	.933	.961	.928	.92	.016	.929	.97	.937	.918	.042	.919	.935	.892	.88	.029	.912	.956	.917	.896	.039	.892	.918	.883	.84

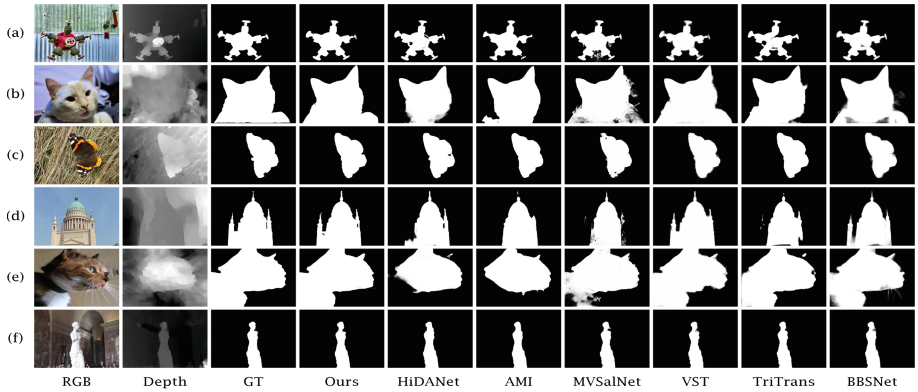


Fig. 5. Visual comparisons with other methods.

ID 2 validates that our designed CSMs is just as effective in the RGB-D SOD task. Compared to ID 0, the values of *Fm* in ID 3 decreases by 0.009 and 0.007 on the two datasets, respectively. In other words, without the help of LPG, the performance of the proposed model deteriorates, which proves that LPG has the ability to guide the model to explore location of salient objects.

Table 2. Effectiveness analysis of general structure.

Method	**ID**	**NJU2K**					**NLPR**				
		***MAE*↓**	***Fm*↑**	***Em*↑**	***Sm*↑**	***Wgt-F*↑**	***MAE*↓**	***Fm*↑**	***Em*↑**	***Sm*↑**	***Wgt-F*↑**
DONet	0	**0.025**	**0.933**	**0.961**	**0.928**	**0.92**	**0.016**	**0.929**	**0.97**	**0.937**	**0.918**
w/o DOIM	1	0.028	0.92	0.955	0.92	0.907	0.017	0.923	0.969	0.933	0.912
w/o CSMs	2	0.028	0.927	0.955	0.922	0.911	0.018	0.921	0.967	0.931	0.908
w/o LPG	3	0.027	0.924	0.956	0.924	0.911	0.017	0.922	0.969	0.935	0.913

Effectiveness of Internal Structures of the Proposed Modules

1. **Effectiveness of MSE and attention mechanisms in DOIM:** The MSE and depth feature optimization are the key components of the proposed DOIM, where the latter is implemented by channel attention and spatial attention. As shown in Table 3, ID 4 and ID 5 represents removing the MSE and the tow attention mechanisms, respectively. It is found that either removing MSE or attention mechanism leads to a decrease in the performance of the proposed model, which proves that both feature enhancement and filtering are indispensable in our DOIM.
2. **Effectiveness of adjacent layer features absorption in CSMs:** In this part, we erase the adjacent layer branches in CSMs and only keep current layer branches

(Table 3, ID 6). The experimental data illustrates that the refinement based solely on the current layer features does not perform as effectively as the additional absorption of features of adjacent layers.

3. **Effectiveness of self-attention mechanism in LPG:** To verify the effectiveness of self-attention in LPG, we remove it in experiment 7, that is, ignoring the transpose, multiplication and channel attention operations. It is observed that the lack of self-attention weakens the location exploration ability of the model, which validates that the proposed self-attention mechanism motivates LPG to explore the location information of salient objects.

Table 3. Effectiveness of internal structures of the proposed modules.

Method	**ID**	**NJU2K**					**NLPR**				
		***MAE*↓**	***Fm*↑**	***Em*↑**	***Sm*↑**	***Wgt-F*↑**	***MAE*↓**	***Fm*↑**	***Em*↑**	***Sm*↑**	***Wgt-F*↑**
DONet	0	**0.025**	**0.933**	**0.961**	**0.928**	**0.92**	**0.016**	**0.929**	**0.97**	**0.937**	**0.918**
w/o MSE	4	0.029	0.925	0.956	0.922	0.91	0.018	0.923	0.968	0.934	0.912
w/o CA,SA	5	0.028	0.927	0.954	0.925	0.913	0.019	0.918	0.964	0.932	0.907
w/current	6	0.027	0.93	0.959	0.926	0.916	0.017	0.924	0.968	0.935	0.912
w/o self-att	7	0.028	0.923	0.955	0.922	0.908	0.019	0.918	0.964	0.929	0.904

5 Conclusions

Following the research trend of multimodal data analysis, we propose a Transformer-based depth optimization network (DONet) for RGB-D SOD task. To maximize the auxiliary effect of depth information, we design a depth feature optimization and integration module (DOIM) retaining high-quality while discarding low-quality depth information. Meanwhile, the proposed context supplement modules (CSMs) refine features by absorbing the features of adjacent layers. For global information exploration, on the one hand, our location perception guider (LPG) utilizes self-attention mechanism to capture the location of salient objects. On the other hand, we employ Transformer as the backbone network, taking advantage of Transformers' ability to comprehensively dig out global information. The proposed DONet achieves significant advantage against 13 state-of-the-art methods on five benchmark datasets.

Acknowledgements. This work is supported in part by grants from the National Natural Science Foundation of China (No. 61806126, 61973307, 61903256, 62062040, 62262030), the Natural Science Foundation of Shanghai (21ZR1462600), the Shanghai Science and Technology Innovation Action Plan (No. 22S31903900), Science and Technology Development Foundation of the Shanghai Institute of Technology (No. ZQ2023-15), the Outstanding Youth Project of Jiangxi Natural Science Foundation (No. 20212ACB212003), the Jiangxi Province Key Subject Academic and Technical Leader Funding Project (No. 20212BCJ23017) and the Natural Science Foundation of Jiangxi Province under Grant (No. 20232BAB202021).

References

1. Chen, H., Li, Y.: Progressively complementarity-aware fusion network for rgb-d salient object detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3051–3060 (2018), 10.1109/CVPR.2018.00322
2. Chen, L.C., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A.L.: Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. IEEE Trans. Pattern Anal. Mach. Intell. **40**(4), 834–848 (2017). https://doi.org/10.1109/TPAMI.2017.2699184
3. Chen, Q., Fu, K., Liu, Z., Chen, G., Du, H., Qiu, B., Shao, L.: Ef-net: A novel enhancement and fusion network for rgb-d saliency detection. Pattern Recogn. **112**, 107740 (2021). https://doi.org/10.1016/j.patcog.2020.107740
4. Chen, Q., Liu, Z., Zhang, Y., Fu, K., Zhao, Q., Du, H.: Rgb-d salient object detection via 3d convolutional neural networks. In: Proceedings of the AAAI conference on artificial intelligence. vol. 35, pp. 1063–1071 (2021), https://doi.org/10.48550/arXiv.2101.10241
5. Chen, S., Fu, Y.: Progressively guided alternate refinement network for rgb-d salient object detection. In: European conference on computer vision. pp. 520–538. Springer (2020), 10.48550/arXiv.2008.07064
6. Chen, Z., Cong, R., Xu, Q., Huang, Q.: Dpanet: Depth potentiality-aware gated attention network for rgb-d salient object detection. IEEE Trans. Image Process. **30**, 7012–7024 (2020). https://doi.org/10.1109/TIP.2020.3028289
7. Cong, R., Lin, Q., Zhang, C., Li, C., Cao, X., Huang, Q., Zhao, Y.: CIR-Net: Cross-modality interaction and refinement for RGB-D salient object detection. IEEE Trans. Image Process. **31**, 6800–6815 (2022). https://doi.org/10.1109/TIP.2022.3216198
8. Cong, R., Liu, H., Zhang, C., Zhang, W., Zheng, F., Song, R., Kwong, S.: Point-aware interaction and cnn-induced refinement network for rgb-d salient object detection. In: Proceedings of the 31st ACM International Conference on Multimedia. pp. 406–416 (2023), https://doi.org/10.1145/3581783.3611982
9. Fan, D.P., Lin, Z., Zhang, Z., Zhu, M., Cheng, M.M.: Rethinking rgb-d salient object detection: Models, data sets, and large-scale benchmarks. IEEE Transactions on neural networks and learning systems **32**(5), 2075–2089 (2020). https://doi.org/10.1109/TNNLS.2020.2996406
10. Fu, K., Fan, D.P., Ji, G.P., Zhao, Q., Shen, J., Zhu, C.: Siamese network for rgb-d salient object detection and beyond. IEEE Trans. Pattern Anal. Mach. Intell. **44**(9), 5541–5559 (2021). https://doi.org/10.1109/TPAMI.2021.3073689
11. Hu, J., Jiang, Q., Cong, R., Gao, W., Shao, F.: Two-branch deep neural network for underwater image enhancement in hsv color space. IEEE Signal Process. Lett. **28**, 2152–2156 (2021). https://doi.org/10.1109/LSP.2021.3099746
12. Ji, W., Li, J., Yu, S., Zhang, M., Piao, Y., Yao, S., Bi, Q., Ma, K., Zheng, Y., Lu, H., et al.: Calibrated rgb-d salient object detection. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 9471–9481 (2021), https://doi.org/10.1109/CVPR46437.2021.00935
13. Jiang, H., Wang, J., Yuan, Z., Wu, Y., Zheng, N., Li, S.: Salient object detection: A discriminative regional feature integration approach. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2083–2090 (2013), https://doi.org/10.1109/CVPR.2013.271
14. Ju, R., Ge, L., Geng, W., Ren, T., Wu, G.: Depth saliency based on anisotropic center-surround difference. In: 2014 IEEE international conference on image processing (ICIP). pp. 1115–1119. IEEE (2014), https://doi.org/10.1109/ICIP.2014.7025222

15. Li, C., Cong, R., Piao, Y., Xu, Q., Loy, C.C.: Rgb-d salient object detection with cross-modality modulation and selection. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VIII 16. pp. 225–241. Springer (2020), https://doi.org/10.48550/arXiv.2007.07051
16. Li, X., Sun, X., Meng, Y., Liang, J., Wu, F., Li, J.: Dice loss for data-imbalanced nlp tasks. arXiv preprint arXiv:1911.02855 (2019), 10.48550/arXiv.1911.02855
17. Liu, N., Zhang, N., Wan, K., Shao, L., Han, J.: Visual saliency transformer. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 4722–4732 (2021), https://doi.org/10.1109/ICCV48922.2021.00468
18. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 10012–10022 (2021), https://doi.org/10.48550/arXiv.2103.14030
19. Liu, Z., Tan, Y., He, Q., Xiao, Y.: Swinnet: Swin transformer drives edge-aware rgb-d and rgb-t salient object detection. IEEE Trans. Circuits Syst. Video Technol. **32**(7), 4486–4497 (2021). https://doi.org/10.1109/TCSVT.2021.3127149
20. Liu, Z., Wang, Y., Tu, Z., Xiao, Y., Tang, B.: Tritransnet: Rgb-d salient object detection with a triplet transformer embedding network. In: Proceedings of the 29th ACM international conference on multimedia. pp. 4481–4490 (2021), https://doi.org/10.1145/3474085.3475601
21. Liu, Z., Zhang, W., Zhao, P.: A cross-modal adaptive gated fusion generative adversarial network for rgb-d salient object detection. Neurocomputing **387**, 210–220 (2020). https://doi.org/10.1016/j.neucom.2020.01.045
22. Niu, Y., Geng, Y., Li, X., Liu, F.: Leveraging stereopsis for saliency analysis. In: 2012 IEEE Conference on Computer Vision and Pattern Recognition. pp. 454–461. IEEE (2012), 10.1109/CVPR.2012.6247708
23. Peng, H., Li, B., Xiong, W., Hu, W., Ji, R.: Rgbd salient object detection: A benchmark and algorithms. In: Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part III 13. pp. 92–109. Springer (2014), https://doi.org/10.1007/978-3-319-10578-9_7
24. Sun, P., Zhang, W., Wang, H., Li, S., Li, X.: Deep rgb-d saliency detection with depth-sensitive attention and automatic multi-modal fusion. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 1407–1417 (2021), https://doi.org/10.1109/CVPR46437.2021.00146
25. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in neural information processing systems **30** (2017), https://doi.org/10.48550/arXiv.1706.03762
26. Wang, F., Pan, J., Xu, S., Tang, J.: Learning discriminative cross-modality features for rgb-d saliency detection. IEEE Trans. Image Process. (2022). https://doi.org/10.1109/TIP.2022.3140606
27. Wang, R., Wang, F., Su, Y., Sun, J., Sun, F., Li, H.: Attention-guided multi-modality interaction network for rgb-d salient object detection. ACM Trans. Multimed. Comput. Commun. Appl. **20**(3), 1–22 (2023). https://doi.org/10.1145/3624747
28. Wang, S., Jiang, F., Xu, B.: Swin transformer-based edge guidance network for rgb-d salient object detection. Sensors **23**(21), 8802 (2023). https://doi.org/10.3390/s23218802
29. Wang, W., Lai, Q., Fu, H., Shen, J., Ling, H., Yang, R.: Salient object detection in the deep learning era: An in-depth survey. IEEE Trans. Pattern Anal. Mach. Intell. **44**(6), 3239–3259 (2021). https://doi.org/10.1109/TPAMI.2021.3051099
30. Wang, W., Xie, E., Li, X., Fan, D.P., Song, K., Liang, D., Lu, T., Luo, P., Shao, L.: Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 568–578 (2021), https://doi.org/10.48550/arXiv.2102.12122

31. Wei, J., Wang, S., Huang, Q.: F^{3}net: fusion, feedback and focus for salient object detection. In: Proceedings of the AAAI conference on artificial intelligence. vol. 34, pp. 12321–12328 (2020), 10.1609/aaai.v34i07.6916
32. Wu, Z., Allibert, G., Meriaudeau, F., Ma, C., Demonceaux, C.: Hidanet: Rgb-d salient object detection via hierarchical depth awareness. IEEE Trans. Image Process. **32**, 2160–2173 (2023). https://doi.org/10.1109/TIP.2023.3263111
33. Yu, J., Jiang, Y., Wang, Z., Cao, Z., Huang, T.: Unitbox: An advanced object detection network. In: Proceedings of the 24th ACM international conference on Multimedia. pp. 516–520 (2016), https://doi.org/10.1145/2964284.2967274
34. Zhai, Y., Fan, D.P., Yang, J., Borji, A., Shao, L., Han, J., Wang, L.: Bifurcated backbone strategy for rgb-d salient object detection. IEEE Trans. Image Process. **30**, 8727–8742 (2021). https://doi.org/10.1109/TIP.2021.3116793
35. Zhang, J., Shi, Y., Yang, J., Guo, Q.: Kd-scfnet: Towards more accurate and lightweight salient object detection via knowledge distillation. Neurocomputing p. 127206 (2023), https://doi.org/10.1016/j.neucom.2023.127206
36. Zhang, M., Yao, S., Hu, B., Piao, Y., Ji, W.: C^{2}dfnet: Criss-cross dynamic filter network for rgb-d salient object detection. IEEE Trans. Multimedia (2022). https://doi.org/10.1109/TMM.2022.3187856
37. Zhao, H., Shi, J., Qi, X., Wang, X., Jia, J.: Pyramid scene parsing network. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2881–2890 (2017), 10.48550/arXiv.1612.01105
38. Zhao, X., Zhang, L., Pang, Y., Lu, H., Zhang, L.: A single stream network for robust and real-time rgb-d salient object detection. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXII 16. pp. 646–662. Springer (2020), https://doi.org/10.48550/arXiv.2007.06811
39. Zhou, J., Wang, L., Lu, H., Huang, K., Shi, X., Liu, B.: Mvsalnet: Multi-view augmentation for rgb-d salient object detection. In: European Conference on Computer Vision. pp. 270–287. Springer (2022), https://doi.org/10.1007/978-3-031-19818-2_16
40. Zhu, C., Li, G.: A three-pathway psychobiological framework of salient object detection using stereoscopic technology. In: Proceedings of the IEEE international conference on computer vision workshops. pp. 3008–3014 (2017), https://doi.org/10.1109/ICCVW.2017.355

Detail-Enhanced Intra- and Inter-modal Interaction for Audio-Visual Emotion Recognition

Tong Shi(✉), Xuri Ge, Joemon M. Jose, Nicolas Pugeault,
and Paul Henderson

School of Computing Science, University of Glasgow, Glasgow, UK
2431206s@student.gla.ac.uk

Abstract. Capturing complex temporal relationships between video and audio modalities is vital for Audio-Visual Emotion Recognition (AVER). However, existing methods lack attention to local details, such as facial state changes between video frames, which can reduce the discriminability of features and thus lower recognition accuracy. In this paper, we propose a Detail-Enhanced Intra- and Inter-modal Interaction network (DE-III) for AVER, incorporating several novel aspects. We introduce optical flow information to enrich video representations with texture details that better capture facial state changes. A fusion module integrates the optical flow estimation with the corresponding video frames to enhance the representation of facial texture variations. We also design attentive intra- and inter-modal feature enhancement modules to further improve the richness and discriminability of video and audio representations. A detailed quantitative evaluation shows that our proposed model outperforms all existing methods on three benchmark datasets for both concrete and continuous emotion recognition. To encourage further research and ensure replicability, our project code is public available at https://github.com/stonewalking/DE-III.

Keywords: Audio-visual emotion recognition · Optical flow · Intra- and Inter-modal modeling · Transformers

1 Introduction

Emotion perception is attracting ever-increasing research attention due to its wide range of applications, such as affective computing [32], human-computer interaction [3], and social robotics [34]. Multi-modal emotion recognition, especially integrating audio and video (i.e. AVER), is particularly important since it makes use of the information present in two modalities that are vital to human communication. Unlike single-modal emotion recognition, multi-modal emotion recognition has access to different representations of the same emotion from different modalities. This improves feature representation capabilities and distinguishability, leading to improved recognition accuracy [8,36].

A. Antonacopoulos et al. (Eds.): ICPR 2024, LNCS 15321, pp. 451–465, 2025.
https://doi.org/10.1007/978-3-031-78305-0_29

However, there are still two challenges that are the focus of ongoing research in AVER: (i) how to enhance the representation of fine details within modalities, such as tiny details of facial motion (e.g. due to micro-expressions), and (ii) how to better leverage inter-modal associations to fully exploit the complementary information from different modalities. Solving both will enable learning better feature representations, and improve emotion recognition accuracy.

When learning features from one modality, intra-modal temporal relationship mining [11,42,44] and feature detail enhancement [39] are important ways to make features more discriminative. For instance, [42] proposed an adaptive graph attention network to explore the relationship between frames of videos for micro-expression recognition, while [39] introduced optical flow to replace face images for micro-expression recognition based on a multi-scale feature representation. However, these methods focus on the single-modal setting, and cannot exploit information from multiple modalities. [44] used self-attention [37] within each modality to enhance their representation and then fused them by a linear-based function to classify; however this cannot fully account for the complex, nonlinear relationships between audio and video.

Multi-modal approaches have recently become mainstream [24,28,41] since considering both audio and video further improves representations, by fusing information in associated video frames and audio fragments. For example, [8] explored the effectiveness of different variants of transformer-based inter-modal attention mechanisms for AVER and showed inter-modal interaction can significantly improve performance. However, although inter-modal interaction improves recognition, these methods do not investigate modeling temporal relationships within each modality. [13] adopts a multi-branch joint auxiliary training method, designing independent audio and video branches and multi-modal fusion to enhance feature relationships, which greatly improves recognition performance. [15,16] used a network shared across modalities to encourage consistency of the multi-modal feature space. However, since different modalities have different feature distributions and properties, a shared network may not fully capture the unique characteristics of each modality, resulting in information loss.

Most of the relationship modeling strategies mentioned above [15–17,35] model temporal relationships based on implicit appearance representation of video frames and audio fragments, but ignore an inherent challenge of AVER – that in video features the frame-to-frame variations of faces are much weaker than in audio. For example, there may be significant changes in content and intonation between two audio fragments, while there is little difference between video frames. It is clear that these missing explicit details, especially state changes between face frames of videos, may lead to reduced discriminability of feature representations during the relationship modeling process, thereby affecting the accuracy of AVER.

We address these issues by introducing a multi-modal interaction network (Figure 1) that incorporates an explicit representation of visual detail changes between frames, and which can better fuse the complementary information from video and audio. Different from methods that directly model relationships between local regions of a facial sequence [10,12,19,30], optical flow is a simple and effective way to represent the state changes between the facial frames. Optical flow can enhance the discriminability of visual representations by directly highlighting significant detail differences between frames, especially those texture changes that can express facial emotions [39]. To this end, we propose a novel detail-enhanced intra- and inter-modal interactions network (called DE-III) for AVER, which integrates explicit optical flow information into an end-to-end multi-modal interaction framework. In addition, two independent multi-modal interaction fusion mechanisms and multiple residual connections further alleviate the information loss problem in existing shared interaction strategies [15,16]. Our main contributions are as follows:

- we explicitly capture detail changes between video frames using optical flow, and integrate this information using a lightweight attentive fusion module;
- we design novel detail-enhanced intra- and inter-modal interaction modules for the video and audio modalities, which can effectively fuse associated information of one modality into the other modality and reduce information loss by residual connections.

We evaluate the resulting model and several variants on three widely used benchmarks and obtain highly competitive results including a new state-of-the-art on multiple metrics, e.g. 83.7% F1-Micro score on CREMA-D, 82.7% accuracy on RAVDESS and the highest scores on MSP-IMPROV with 89.3%, 88.7% and 85.8% for valence, arousal and dominance.

2 Related Work

Emotion recognition has received a significant amount of attention in the computer vision community. Numerous methods [22,31,38,43] have been proposed to solve this task by using different data modalities, such as images, speech and text. These methods can be divided into two main kinds: unimodal methods (that input just one modality), and multi-modal methods (that input two or more modalities). Our proposed DE-III belongs to the latter category, combining audio and video modalities to improve the performance of emotion recognition.

2.1 Unimodal Emotion Recognition

Unimodal emotion recognition methods [1,22,31,38,43] focus on application scenarios where only one kind of data is available; they design feature enhancement and interaction methods based on the inherent properties of the corresponding modality. The most common methods are text-based [1,9,40] and image-based [31,38,43]. For example for text, [1] present a BERT-based model to explore the importance of context extraction in texts for emotion recognition. One work by [33] proposed one sequence-based convolutional neural network to detect human emotion from big data. However, it is harder to to accurately predict human emotions from a text transcription compared to using richer modalities such as images or videos. For image data, [31] proposed feature decomposition and reconstruction learning for effective facial image expression recognition. [27] introduced the image depth information to improve the context information of images, which improved the representation capability and thus recognition accuracy. Moving to video, [2] introduced facial micro-expression analysis methods that can improve emotion recognition by capturing richer contextual sequence information than static images. Although unimodal emotion recognition has achieved substantial progress and delivers promising results, it is inherently limited by having less information available than multi-modal approaches.

2.2 Multi-modal Emotion Recognition

Recently, multi-modal emotion recognition has become mainstream [7,8,13–17,26,36] due to its ability to fully exploit the complementary information present in different modalities. For instance, [8] explored the effectiveness of different variants of transformer-based inter-modal attention mechanisms for audio-video emotion recognition and showed that inter-modal interaction can significantly improve performance. [25] showed that combining audio and with a corresponding text transcription improves the representation ability of features, since audio captures details of intonation, while text captures semantics more explicitly. Moreover, [7] fused three modalities (audio, text and vision), further improving recognition accuracy. The above works indicate that combining multiple modalities can significantly enhance the discrimination ability of fused representations and thus the recognition performance. In this work, we study multimodal-based emotion recognition, specifically for audio-video emotion recognition (AVER). The most similar works to ours are [15,16], both of which used a transformer-based architecture that is shared across video and audio modalities to encourage consistency of the multi-modal feature space. However, their proposed shared network cannot fully capture the unique feature distributions of each modality, such as explicit facial state changes between face video frames, resulting in the loss of information during the multimodal relationship modeling process. Unlike [13,14,36], which adopt attention-based neural network to effectively process and integrate audio modalities, our model not only learns the intra-relationships within video feature representations but also models the inter-relationships when attentively fuses the audio representation. Our proposed model augments video features with optical flow information before fusing with the audio features. Unlike traditional methods [21] that directly combine the optical flow features with visual representations, we use Conformer [17] networks to extract context-aware features, and design a novel pairwise O-V attention fusion module to combine them.

3 Proposed Model

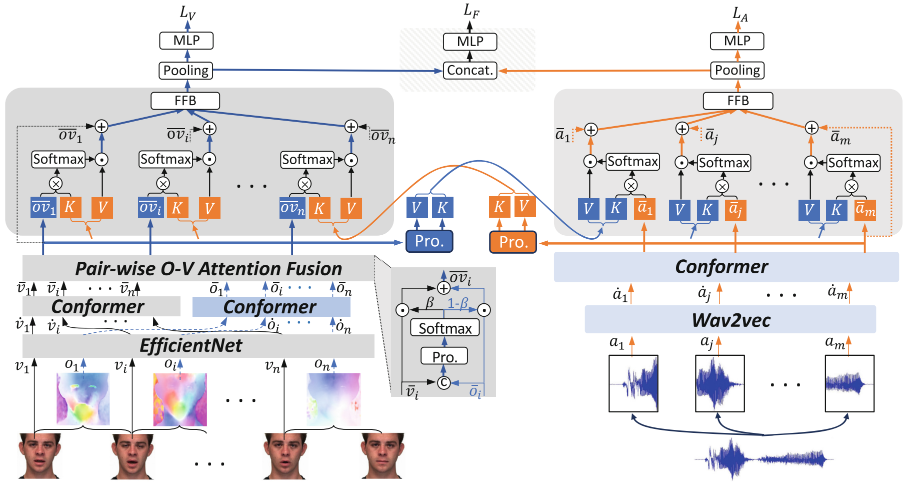


Fig. 1. Overview of our proposed method *DE-III*. Given video frames v_i and audio fragments a_i, we extract features and pass these through separate Conformer encoders. We introduce explicit information about facial motions – captured by optical flow o_i – to enhance video feature representations, with a new pair-wise O-V attention fusion module that effectively integrates the information from optical flow and video frames. We propose an inter-modal feature enhancement module (large boxes near top) to attentively fuse the associated audio and video representations in both directions, i.e. audio-to-video and video-to-audio. During training, the final emotion predictions are calculated independently from three sets of features: the video features albeit with audio information fused (i.e. without the model components in the chequered box); the converse using the audio features; and finally using both sets of features after a further fusion stage. During inference, we use the prediction head that performed best on validation data.

The overall framework of our proposed model DE-III is shown in Figure 1. We first extract video and audio features, then enhance their representative power through temporal relationship modelling within their respective modalities, also fusing optical flow information with the video features to better capture detail changes. Then, the inter-modal feature enhancement module performs attention-weighted fusion of each modality's information with the other modality.

3.1 Audio Self-enhancement Module

To represent the information in audio, we use a pre-trained wav2vec model [18] to embed the extracted audio fragments.The original speech audio is resampled at 16 kHz. Specifically, we split a given audio clip into a sequence of m fragments $A = \{a_1, a_2, \ldots, a_m\}$ using a sliding window. Then we use the wav2vec-large-robust model to extract corresponding fragment-level representations $\dot{A} = \{\dot{a}_1, \dot{a}_2, \ldots, \dot{a}_m\}$. Next, a Conformer encoder [17] (a transformer-based model with convolutions to improve temporally-local information processing) is used to obtain enhanced audio-fragment representations $\bar{A} = \{\bar{a}_1, \bar{a}_2, \ldots, \bar{a}_m\}$ that account for (intra-modal) local and global temporal relationships.

3.2 Video Pairwise Attention Enhancement Module

Different from the audio features where contextual semantics are clear, i.e. there is clear semantic content and significant intonation changes, in the video, subtle yet important changes in facial texture tend to be lost during feature extraction. We therefore use a pre-trained optical flow model [20] to extract the flow o_i between adjacent pairs of video frames $\{v_{i-1}, v_i\}$, where $i \in \{1, \ldots, n\}$ and n is the number of video frames; this can explicitly represent fine-grained changes of facial texture such as micro-expressions. Then, we employ the widely-used EfficientNet-B2 model [32], which has been fine-tuned on VGGface2 [5] dataset, to extract representations for video frames and their corresponding optical flow maps; we denote these features by $\dot{V} = \{\dot{v_1}, \dot{v_2}, \ldots, \dot{v_n}\}$ and $\dot{O} = \{\dot{o_1}, \dot{o_2}, \ldots, \dot{o_n}\}$ respectively. To further enhance the representational ability of these visual features, we use two independent Conformer encoders [17] to embed them into the same dimensional space as the audio modality. This also allows for subsequent inter-modal interaction. We next propose a simple and efficient pairwise O-V attention fusion module to combine the features of frames and optical flow into a joint embedding space. Specifically, we use a fully-connected (FC) layer to map the features at each time-point to two channels, then apply a softmax function [6] and interpret these values as weights for the frame and flow features respectively. We finally obtain the detailed-enhanced video representation $\overline{ov}_i$ by a weighted sum of linearly-projected frame features and corresponding flow features. Thus, we set

$$[\bar{o}_i : \bar{v}_i] = [\text{Conformer}(\dot{o}_i) : \text{Conformer}(\dot{v}_i)], \tag{1}$$

$$(\beta_o, \beta_v) = \text{softmax}(FC([\bar{o}_i : \bar{v}_i])), \tag{2}$$

$$\overline{ov}_i = \beta_o W_o \bar{o}_i + \beta_v W_v \bar{o}_i, \tag{3}$$

where $[:]$ denotes concatenation along the channel dimension, W_o and W_v are the linear projection parameters, and $\beta_o + \beta_v = 1$. We refer to the two conformers followed by the OV-fusion as the pair-wise attention enhancement (PAE) module.

3.3 Inter-modal Feature Enhancement Module

Inspired by the attention mechanisms [13,41], we next design an inter-modal feature enhancement module (IFE) that allows each modality to attend to the other and integrate relevant information. For simplicity we describe only the audio-to-video fusion (IFE-Video); however a similar approach is used for video-to-audio. We want to allow the enhanced video frame features $\overline{ov}_i$ to attend to features of relevant audio fragments $\bar{A} = \{\bar{a}_1, \bar{a}_2, \ldots, \bar{a}_m\}$. Different from traditional self-attention [37] and cross-attention [36], we take the target video frame $\overline{ov}_i$ as the query to calculate the attention weights, with the audio fragment defining the keys and values after the linear projections. Attentive fusion from another modality allows relevant modality information to be extracted and integrated, thereby improving the distinguishability of target modality representation. Finally, we obtain the video representations $\ddot{OV} = \{\ddot{ov}_i\}$ after IFE by adding a residual connection, and passing through a feed-forward block (FFB) which contains two linear layers. In summary, we set

$$s_{ij} = \frac{(W_{ov}\overline{ov}_i)(W_a\bar{a}_j)^T}{||W_{ov}\overline{ov}_i||\ ||W_a\bar{a}_j||} \quad \forall\, i \in \{1, \ldots, n\},\, j \in \{1, \ldots, m\} \tag{4}$$

$$\alpha_{ij} = \frac{\exp(s_{ij})}{\sum_{j=1}^{m}\exp(s_{ij})} \tag{5}$$

$$\ddot{ov}_i = \sum_{j=1}^{m} \alpha_{ij}\bar{W}_a\bar{a}_j + \overline{ov}_i, \tag{6}$$

where W_{ov}, W_a and $\bar{W}_a$ are linear projection parameters. Similarly, we obtain the attention-aware video fragment representations of each audio fragment and combine them with an audio residual operation to give the final audio representations $\ddot{A} = \{\ddot{a}_1, \ddot{a}_2, ..., \ddot{a}_m\}$.

3.4 Feature Aggregation and Objective Function

Since we want to make a single prediction for an entire video, we max-pool the features along the temporal axis, yielding a video-centric feature vector $\ddot{ov}^*$ from $\ddot{OV}$, and audio-centric feature vector $\ddot{a}^*$ from $\ddot{A}$ (note that $\ddot{ov}^*$ still incorporates information fused from the audio modality as described in Section 3.3, and vice-versa). We use three independent emotion prediction heads (each a multi-layer perceptron) with corresponding losses to jointly optimize different branches the model – the video-cross loss L_V (using $\ddot{ov}^*$ as input to the MLP), audio-cross loss L_A (using $\ddot{a}^*$) and audio-visual fusion loss L_F (using $\ddot{ov}^*$ concatenated with $\ddot{a}^*$). The overall objective function is the sum of the three losses. We use multi-class cross-entropy for datasets with discrete emotion class labels, and concordance correlation coefficient (CCC) for datasets with continuous labels. Specifically, CCC is given by

$$\mathcal{L}_{\text{CCC}} = 1 - \frac{2\rho\sigma_x\sigma_y}{\sigma_x^2 + \sigma_y^2 + (\mu_x - \mu_y)^2} \tag{7}$$

where μ_x and μ_y are the mean of the predicted result $\hat{y}$ and the label y, respectively, σ_x and σ_y are their standard deviations, and ρ is their Pearson correlation coefficient (a ρ value close to ±1 suggests a strong linear relationship, while a value of 0 signifies the absence of any linear correlation). During inference we can use predictions from any of the three heads; for our main experiments we use the prediction head that performed best on the validation data.

4 Experiments

4.1 Experimental Setup

Datasets and Metrics. To verify the effectiveness of our proposed approach, we evaluate it on three popular AVER datasets: CREMA-D [4], MSP-IMPROV [3] and RAVDESS [23]. CREMA-D consists of 7,442 facial videos with corresponding audio from 96 participants (48 male, 48 female). Each audio-video clip is labeled with one of 6 concrete emotion classes – anger, disgust, fear, happiness, sadness, and neutrality. RAVDESS consists of 2,880 videos from 24 actors, each enacting eight concrete emotional states. MSP-IMPROV consists of 8,385 audio-video clips from 12 participants (6 male, 6 female) with each clip labeled by both concrete emotional states and continuous emotional states – valence, arousal and dominance; following previous works [8,14,16,36] we use only the continuous labels. We adhered to the protocol in [8,13], with 5 separate folds where each fold divides the data into training, validation, and test sets with non-overlapping actor identities. We evaluate based on the most commonly-used metrics for each dataset – F1-Macro and F1-Micro for CREMA-D [4], Accuracy for RAVDESS [23] and CCC for MSP-IMPROV [3].

Table 1. Comparisons with state-of-the-art methods for AVER on CREMA-D, MSP-IMPROV and RAVDESS (in %). The best results are bold and second-best underlined.

Method	CREMA-D		MSP-IMPROV			RAVDESS
	F1-Macro	F1-Micro	Val.	Aro.	Dom.	Acc.
Multi. [36]	64.4	69.2	77.5	76.1	77.8	78.5
MMER [8]	–	–	–	–	–	81.6
UAVM [16]	74.9	76.9	47.1	54.4	68.7	–
AuxFormer [13]	69.8	76.3	67.2	65.2	82.0	–
LADDER [14]	**80.2**	80.3	–	–	–	–
DE-III (ours)	79.5	**83.7**	**89.3**	**88.7**	**85.8**	**82.7**

Implementation Details. All models were trained for up to 20 epochs using early stopping on the validation set, and we report our results on the test set. We choose hyper-parameters based on validation set performance. We use AdamW for optimization with a learning rate of 5×10^{-6} and weight decay of 5×10^{-2}. The face images are extracted from each frame of every video clip and resized to 224×224 pixels. We generate optical flow maps using [20] and normalize their magnitude by a standard deviation calculated from the local optical flow magnitude at every pixel position within an entire video clip. We use the pre-trained EfficientNet-B2 from [32] to extract features from the video frames and optical

Table 2. Effectiveness of our inter-modal feature enhancement module (IFE), evaluated on CREMA-D.

Method	Cross attention		Accuracy	
	A-cross	V-cross	F1-Macro	F1-Micro
IFE-Fusion	✓	✓	77.2	82.2
IFE-Audio	✓	✗	78.3	82.2
IFE-Video	✗	✓	**79.5**	**83.7**
None-IFE	✗	✗	75.8	78.6

flow maps. The audio features are extracted using wav2vec2-large-robust [18]. Separate Conformer encoders for video and audio map the extracted features to vectors of 1408-dimension each. Each Conformer block has a hidden dimensionality of 512, with 8 attention heads. The number of blocks in the acoustic, visual, and optical flow Conformers were set to 3, 3, and 2, respectively. For the prediction heads, we use MLPs with hidden dimensionality of 512. Our IFE module (Section 3.3) uses single-head attention [37] with the linear feed-forward block and the highlighted fusion feature dimensions remain unchanged. Our model was implemented in PyTorch and trained on 2 NVIDIA RTX A5000 GPUs, taking 1 hour.

4.2 Quantitative Comparison

In Table 1 we present quantitative results for our method and several existing works: 1) Multi [36], a transformer-based cross-modal attention fusion method; 2) MMER [8], with multiple self-attention fusion mechanisms; 3) UAVM [16], a transformer-based feature enhancement model with a shared audio-visual encoder; 4) AuxFormer [13], a transformer framework with two independent auxiliary branches; 5) LADDER [14], a transformer-based cross-attention framework with auxiliary reconstruction tasks. We see that compared with the previous best method LADDER [14] on CREMA-D, our DE-III achieves higher performance in terms of F1-Micro score, 83.7% vs. 80.3%. On MSP-IMPROV, our DE-III attains excellent CCC values of 89.3% for valence (Val.), 88.7% for arousal (Aro.), and 85.8% for dominance (Dom.), establishing a new state-of-the-art for this dataset. Moreover, we also achieve a better accuracy (Acc.) score on RAVDESS compared with the SOTA method, 82.7% for DE-III vs. 81.6% for MMER.

4.3 Ablation Studies

In this section, we evaluate the performance benefit due to various components and design decisions in our model.

Effects of Inter-modal Feature Enhancement (IFE). In the IFE block, we define video attending to audio as V-cross and audio attending to video as

A-cross. We first experiment with removing the IFE module (i.e. without any inter-modality fusion, only RGB images and flow maps, denoted None-IFE). In Table 2, we see a large performance drop in this setting – compared with the best output (from IFE-Video), the F1-Macro and F1-Micro scores decrease by 3.7% and 5.1% on the CREMA-D test set, respectively. This suggests that inter-modality fusion plays an important role in improving AVER capabilities. Recall that our model has three prediction heads: IFE-Audio (i.e. using features $\ddot{a}^*$), IFE-Video (i.e. using $\ddot{o}\ddot{v}^*$) and IFE-Fusion (i.e. using their concatenation). While the main results use IFE-Video at inference time, we also report results from the others in Table 2. IFE-Video achieves the best AVER performance, 79.5% F1-Macro and 83.7% F1-Micro. The other prediction heads achieve slightly lower though still competitive results.

Table 3. Effectiveness of different approaches to inter-modal fusion within our model, evaluated on CREMA-D.

Model	Fuse when?		Visual input		Seq. model		Fuse how?			Accuracy	
	Early	Late	Flow	RGB	Conf.	Transf.	Concat	Sum	PAE	F1-Macro	F1-Micro
IFE-V-O	n/a		✓		✓		n/a			55.4	64.9
IFE-V-F	n/a			✓	✓		n/a			76.7	81.4
IFE-V-FOSC		✓	✓	✓	✓		✓			78.5	81.7
IFE-V-FODC		✓	✓	✓	✓		✓			77.8	82.6
IFE-V-FODS		✓	✓	✓	✓			✓		78.0	81.8
IFE-V-Early	✓		✓	✓	✓				✓	79.2	83.0
IFE-V-Trans		✓	✓	✓		✓			✓	77.9	82.6
IFE-Video		✓	✓	✓	✓				✓	**79.5**	**83.7**

Effects of Video Pairwise Attention Enhancement (PAE) Module. To demonstrate our ablations on pair-wise attention enhancement (PAE) Module, we categorize different settings as "Fuse when?", "Visual input", "sequential model", and "Fuse how?". Results on CREMA-D are given in Table 3, all using the IFE-Video prediction head. We first present results when trained with only one part of the video information, i.e. RGB images only (IFE-V-F), or optical flow maps only (IFE-V-O). We see that IFE-V-O achieves 55.4% F1-macro and 64.9% F1-micro. The result shows optical flow information present low capability to distinguish emotions, and it is much weaker than using RGB images only. When combining optical flow maps with RGB images in the full model (IFE-Video), there is a remarkable performance improvement vs. IFE-V-F. It indicates that the flow maps indeed augment the video feature representations. Next, we replace our PAE with one single conformer followed by one OV-fusion block. To pass the image and optical flow features together into the conformer, we attempt several alternative operations– temporal concatenation (IFE-V-FOSC), channelwise concatenation (IFE-V-FODC), and summation (IFE-V-FODS). We see (Table 3) that our PAE module achieves the highest recognition performance, with 1.0% improvement over IFE-V-FOSC on F1-macro and 1.1% improvement over IFE-V-FODC on F1-Micro. These observations indicate that our PAE module is a more effective fusion method for combining visual features and optical flow features. Finally, we explore early fusion and late fusion strategies. We find that by moving OV-fusion block before the Conformer (IFE-V-Early), accuracy decreases slightly vs. having OV-fusion after the Conformer (IFE-Video), by 0.3% F1-Macro and 0.7% F1-Micro. We hypothesise that this is because the additional computation performed beforehand by the Conformer is beneficial in helping the OV-fusion module to determine whether to focus on image or flow information for each time-point. Additionally, we compare our method by replacing the conformer to the vanilla transformer [37], the accuracy decreases slightly by 1.6% and 1.1%, this demonstrates that the conformer is superior to the vanilla transformer at the image level in capturing changes in facial details from feature representations.

Table 4. Effectiveness of different feature extractors and frame-selection strategies for optical-flow, evaluated on CREMA-D for our IFE-Video model variant.

Feature extractor	Window	Stride	Accuracy	
			F1-Macro	F1-Micro
EfficientNet-B2 [32]	1	1	**79.5**	**83.7**
	3	1	76.1	81.4
	5	1	74.3	80.8
	7	1	75.2	81.6
	3	3	77.2	82.4
	5	5	76.2	80.7
	7	7	78.5	82.8
DINOv2 [29]	1	1	76.8	82.6

Effects of optical-flow extraction variants. We next experiment with using different sliding window lengths and strides when extracting the optical flow from the videos. Firstly, we vary the window length while keeping the stride fixed to 1 (i.e. moving frame by frame). Secondly, we vary both the window length and the stride together (i.e. non-overlapping windows). The results in Table 4 show that using a window length of 1 with a stride of 1 performs best. Increased window lengths, with fixed or increasing strides, show consistent drops in performance, with the worst-performing variant having window length of 5 and stride of 1 (achieving 74.3% F1-Macro, versus 79.5% for window length and stride of 1). This indicates that temporally-fine-grained information is valuable in increasing the accuracy of emotion recognition. We also experiment with using a different backbone feature extractor for the optical flow, since face images and flow-maps

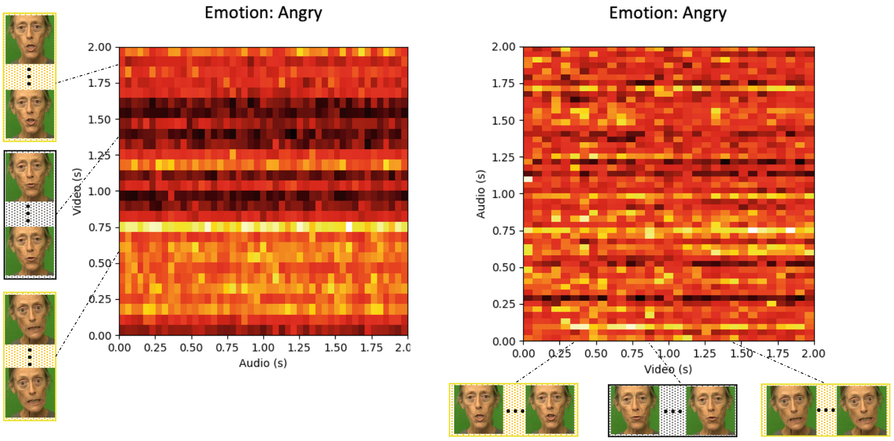


Fig. 2. Heatmaps showing inter-modality attention weights calculated by IFE-Audio (left) and IFE-Video (right), for an example sequence with emotion 'angry'. The horizontal axis corresponds to time-points in one modality, which is fusing in information from the other modality on the vertical axis. Brighter colors indicate stronger attention to the time-point on the vertical axis, from the time-point on the horizontal axis.

are quite different domains. We choose DINOv2 [29], which has been shown to be robust across many image domains, and fix the window length and stride to 1 (i.e. the best-performing setting). However, we find it performs worse than using EffcientNet pre-trained on a large face images dataset, dropping from 79.5% to 76.8% F1-Macro and from 83.7% to 82.6% F1-Micro.

4.4 Qualitative Analysis

To better understand the behavior of our model, we visualize the inter-modal fusion weights α_{ij} for IFE-Audio and IFE-Video (see Section 3.3) in Figure 2. The brightness of each location in the heatmap represents the strength with which the modality on the horizontal axis is attending to that on the vertical axis, at that particular time-point. The pattern of attention varies considerably for different points along the horizontal axis, showing that the model does not attend to fixed, specific points in the other modality, but adapts depending on the current features, and presumably the varying emotional states depicted in the video. Notably, the heatmaps do not exhibit a bright diagonal line; this indicates that time-points generally attend not to the corresponding time-point in the other modality, but to other (presumably relevant or informative) time-points. Overall these results suggest that our inter-modal feature enhancement module can selectively fuse the useful information from each modality into the other.

5 Conclusion

We have presented a new model, DE-III, for audio-visual emotion recognition, which combines intra- and inter-model feature enhancement in a unified framework. DE-III introduces a pair-wise attention fusion method that integrates explicit facial detail changes between video frames, captured by optical flow. It not only improves the distinguishability of features within each visual modality, but also further increases the effectiveness of subsequent inter-modal feature interactions. Our results demonstrate that DE-III enhances emotion recognition by optimally fusing the information available in different modalities. Indeed, our model achieves state-of-the-art performance on three popular datasets, for both concrete and continuous emotion labels.

References

1. Acheampong, F.A., Nunoo-Mensah, H., Chen, W.: Transformer models for text-based emotion detection: a review of bert-based approaches. Artif. Intell. Rev. **54**(8), 5789–5829 (2021)
2. Ben, X., Ren, Y., Zhang, J., Wang, S.J., Kpalma, K., Meng, W., Liu, Y.J.: Video-based facial micro-expression analysis: A survey of datasets, features and algorithms. IEEE Trans. Pattern Anal. Mach. Intell. **44**(9), 5826–5846 (2021)
3. Busso, C., Parthasarathy, S., Burmania, A., AbdelWahab, M., Sadoughi, N., Provost, E.M.: MSP-IMPROV: An acted corpus of dyadic interactions to study emotion perception. Trans. Affect. Comput. **8**(1), 67–80 (2016)
4. Cao, H., Cooper, D.G., Keutmann, M.K., Gur, R.C., Nenkova, A., Verma, R.: CREMA-D: Crowd-sourced emotional multimodal actors dataset. Trans. Affect. Comput. **5**(4), 377–390 (2014)
5. Cao, Q., Shen, L., Xie, W., Parkhi, O.M., Zisserman, A.: VGGFace2: A dataset for recognising faces across pose and age. In: International Conference on Automatic Face and Gesture Recognition (2018)
6. Chorowski, J.K., Bahdanau, D., Serdyuk, D., Cho, K., Bengio, Y.: Attention-based models for speech recognition. NeurIPS **28** (2015)
7. Chudasama, V., Kar, P., Gudmalwar, A., Shah, N., Wasnik, P., Onoe, N.: M2fnet: Multi-modal fusion network for emotion recognition in conversation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4652–4661 (2022)
8. Chumachenko, K., Iosifidis, A., Gabbouj, M.: Self-attention fusion for audiovisual emotion recognition with incomplete data. In: ICPR. pp. 2822–2828. IEEE (2022)
9. Deng, J., Ren, F.: A survey of textual emotion recognition and its challenges. IEEE Trans. Affect. Comput. **14**(1), 49–67 (2021)
10. Ge, X., Jose, J.M., Wang, P., Iyer, A., Liu, X., Han, H.: Algrnet: Multi-relational adaptive facial action unit modelling for face representation and relevant recognitions. Behavior, and Identity Science, IEEE Transactions on Biometrics (2023)
11. Ge, X., Jose, J.M., Xu, S., Liu, X., Han, H.: Mgrr-net: Multi-level graph relational reasoning network for facial action unit detection. ACM Transactions on Intelligent Systems and Technology **15**(3), 1–20 (2024)
12. Ge, X., Wan, P., Han, H., Jose, J.M., Ji, Z., Wu, Z., Liu, X.: Local global relational network for facial action units recognition. In: FG. pp. 01–08. IEEE (2021)

13. Goncalves, L., Busso, C.: AuxFormer: Robust approach to audiovisual emotion recognition. In: ICASSP. pp. 7357–7361. IEEE (2022)
14. Goncalves, L., Busso, C.: Learning cross-modal audiovisual representations with ladder networks for emotion recognition. In: ICASSP. pp. 1–5. IEEE (2023)
15. Goncalves, L., Leem, S.G., Lin, W.C., Sisman, B., Busso, C.: Versatile audio-visual learning for handling single and multi modalities in emotion regression and classification tasks. arXiv preprint arXiv:2305.07216 (2023)
16. Gong, Y., Liu, A.H., Rouditchenko, A., Glass, J.: Uavm: Towards unifying audio and visual models. IEEE Signal Process. Lett. **29**, 2437–2441 (2022)
17. Gulati, A., Qin, J., Chiu, C.C., Parmar, N., Zhang, Y., Yu, J., Han, W., Wang, S., Zhang, Z., Wu, Y., et al.: Conformer: Convolution-augmented transformer for speech recognition. In: ICASSP. p. 5749–5753. IEEE (2021)
18. Hsu, W.N., Sriram, A., Baevski, A., Likhomanenko, T., Xu, Q., Pratap, V., Kahn, J., Lee, A., Collobert, R., Synnaeve, G., et al.: Robust wav2vec 2.0: Analyzing domain shift in self-supervised pre-training. arXiv preprint arXiv:2104.01027 (2021)
19. Hu, M., Ge, P., Wang, X., Lin, H., Ren, F.: A spatio-temporal integrated model based on local and global features for video expression recognition. The Visual Computer pp. 1–18 (2021)
20. Jaegle, A., Borgeaud, S., Alayrac, J.B., Doersch, C., Ionescu, C., Ding, D., Koppula, S., Zoran, D., Brock, A., Shelhamer, E., et al.: Perceiver IO: A general architecture for structured inputs & outputs. arXiv preprint arXiv:2107.14795 (2021)
21. Kemmou, A., El Makrani, A., El Azami, I., Aabidi, M.H.: Automatic facial expression recognition under partial occlusion based on motion reconstruction using a denoising autoencoder. Indonesian Journal of Electrical Engineering and Computer Science **34**(1), 276–289 (2024)
22. Li, X., Zhang, Y., Tiwari, P., Song, D., Hu, B., Yang, M., Zhao, Z., Kumar, N., Marttinen, P.: Eeg based emotion recognition: A tutorial and review. ACM Comput. Surv. **55**(4), 1–57 (2022)
23. Livingstone, S.R., Russo, F.A.: The Ryerson audio-visual database of emotional speech and song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English. PLoS ONE **13**(5), e0196391 (2018)
24. Ma, F., Zhang, W., Li, Y., Huang, S.L., Zhang, L.: An end-to-end learning approach for multimodal emotion recognition: Extracting common and private information. In: ICME. pp. 1144–1149 (2019)
25. Maji, B., Swain, M., Guha, R., Routray, A.: Multimodal emotion recognition based on deep temporal features using cross-modal transformer and self-attention. In: ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 1–5. IEEE (2023)
26. Mittal, T., Bhattacharya, U., Chandra, R., Bera, A., Manocha, D.: M3er: Multiplicative multimodal emotion recognition using facial, textual, and speech cues. In: Proceedings of the AAAI conference on artificial intelligence. vol. 34, pp. 1359–1367 (2020)
27. Mittal, T., Guhan, P., Bhattacharya, U., Chandra, R., Bera, A., Manocha, D.: Emoticon: Context-aware multimodal emotion recognition using frege's principle. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14234–14243 (2020)
28. Nie, W., Ren, M., Nie, J., Zhao, S.: C-GCN: Correlation based graph convolutional network for audio-video emotion recognition. IEEE Trans. Multimedia **23**, 3793–3804 (2020)

29. Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., et al.: Dinov2: Learning robust visual features without supervision. arXiv preprint arXiv:2304.07193 (2023)
30. Perveen, N., Roy, D., Chalavadi, K.M.: Facial expression recognition in videos using dynamic kernels. Trans. Image Process. **29**, 8316–8325 (2020)
31. Ruan, D., Yan, Y., Lai, S., Chai, Z., Shen, C., Wang, H.: Feature decomposition and reconstruction learning for effective facial expression recognition. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 7660–7669 (2021)
32. Savchenko, A.V.: Facial expression and attributes recognition based on multi-task learning of lightweight neural networks. In: SISY. pp. 119–124. IEEE (2021)
33. Shrivastava, K., Kumar, S., Jain, D.K.: An effective approach for emotion detection in multimedia text data using sequence based convolutional neural network. Multimedia tools and applications **78**, 29607–29639 (2019)
34. Spezialetti, M., Placidi, G., Rossi, S.: Emotion recognition for human-robot interaction: Recent advances and future perspectives. Frontiers in Robotics and AI **7** (2020). 10.3389/frobt.2020.532279
35. Tarantino, L., Garner, P.N., Lazaridis, A., et al.: Self-attention for speech emotion recognition. In: Interspeech. pp. 2578–2582 (2019)
36. Tsai, Y.H.H., Bai, S., Liang, P.P., Kolter, J.Z., Morency, L.P., Salakhutdinov, R.: Multimodal transformer for unaligned multimodal language sequences. In: ACL. vol. 2019, p. 6558. NIH Public Access (2019)
37. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. NeurIPS **30** (2017)
38. Wang, K., Peng, X., Yang, J., Lu, S., Qiao, Y.: Suppressing uncertainties for large-scale facial expression recognition. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 6897–6906 (2020)
39. Wang, M.: Micro-expression recognition based on multi-scale attention fusion. In: ICDSCA. pp. 853–861. IEEE (2021)
40. Wang, X., Kou, L., Sugumaran, V., Luo, X., Zhang, H.: Emotion correlation mining through deep learning models on natural language text. IEEE transactions on cybernetics **51**(9), 4400–4413 (2020)
41. Yin, Y., Jing, L., Huang, F., Yang, G., Wang, Z.: Msa-gcn: Multiscale adaptive graph convolution network for gait emotion recognition. Pattern Recognition p. 110117 (2023)
42. Zhang, Y., Wang, H., Xu, Y., Mao, X., Xu, T., Zhao, S., Chen, E.: Adaptive graph attention network with temporal fusion for micro-expressions recognition. In: ICME. pp. 1391–1396. IEEE (2023)
43. Zhang, Y., Wang, C., Deng, W.: Relative uncertainty learning for facial expression recognition. Adv. Neural. Inf. Process. Syst. **34**, 17616–17627 (2021)
44. Zhou, H., Meng, D., Zhang, Y., Peng, X., Du, J., Wang, K., Qiao, Y.: Exploring emotion features and fusion strategies for audio-video emotion recognition. In: ICMI. pp. 562–566 (2019)

Author Index

A. Antonacopoulos et al. (Eds.): ICPR 2024, LNCS 15321, pp. 467–468, 2025.
https://doi.org/10.1007/978-3-031-78305-0