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

LNCS 15332

Pattern Recognition

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

32 Part XXXII

ICPR 2024 INDIA

IAPR

Springer

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 XXXII

Springer

Editors
Apostolos Antonacopoulos
University of Salford
Salford, Lancashire, UK

Rama Chellappa
Johns Hopkins University
Baltimore, MD, USA

Saumik Bhattacharya
IIT Kharagpur
Kharagpur, West Bengal, India

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

Cheng-Lin Liu
Chinese Academy of Sciences
Beijing, China

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

ISSN 0302-9743 ISSN 1611-3349 (electronic)
Lecture Notes in Computer Science
ISBN 978-3-031-78124-7 ISBN 978-3-031-78125-4 (eBook)
https://doi.org/10.1007/978-3-031-78125-4

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 biosignal 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. In ICRP 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 Institute of Technology, 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	Indian Institute of Technology, 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 XXXII

Pixel Embedding for Fractional Interpolation in Video Coding

Young-Woon Lee[1], Qurat Ul Ain Aisha[2], and Byung-Gyu Kim[2](✉)

[1] Sunmoon University, Asan, Korea
{yw.lee, aisha.q}@ivpl.sm.ac.kr
[2] Sookmyung Women's University, Seoul, Korea
bg.kim@sookmyung.ac.kr

Abstract. In July 2020, Versatile Video Coding (VVC/H.266) has been finalized as the next-generation video coding standard. Due to the diverse characteristics of video, motion prediction in fractional precision is required in the video coding. For this, VVC/H.266 uses Discrete Cosine Transform-based Interpolation Filter (DCTIF) but, it is being a typical low-pass filter with fixed integer coefficients so it cannot guarantee optimal performance across all videos. Recently, deep learning-based technology has been continually developed onwards. This paper proposed the In-Loop Interpolation Filter (ILIF) which can generate high-quality fractional pixels. ILIF is an Super-Resolution (SR) model with our proposed pixel embedding technique. Pixel Embedding allows the correlation between integer and sub-pixels to be maintained during learning and it is highly effective in the inter coding. Optimized through a divide-and-conquer learning approach, ILIF replaces the DCTIF and is integrated with inter prediction in VVC/H.266. ILIF considered only the Y component of YUV420 format and the BD-rate performance was compared and analyzed with the anchor of VVC/H.266. Two integration methods (MODE 1, 2) between ILIF and VVC/H.266 were presented. As a result of the experiment, for MODE 1 which applies ILIF only for fractional pixel generation, the gains were −1.42% for All-QP, −1.54% for High-QP, and −1.24% for Low-QP. Additionally, in MODE 2 which integrates integer pixel filtering and sub-pixel generation with ILIF, it showed the gains of −3.92% for All-QP, −4.01% for High-QP, and −3.13% for Low-QP.

Keywords: Versatile Video Coding (VVC/H.266) · fractional interpolation · Convolutional Neural Network (CNN) · Pixel Embedding

1 Introduction

In July 2020, Versatile Video Coding (VVC/H.266) was officially announced as the final standard [2]. VVC/H.266 was developed with the goal of achieving

Y.-W. Lee and Q. Ul Ain Aisha—Equal contribution.

A. Antonacopoulos et al. (Eds.): ICPR 2024, LNCS 15332, pp. 1–16, 2025.
https://doi.org/10.1007/978-3-031-78125-4_1

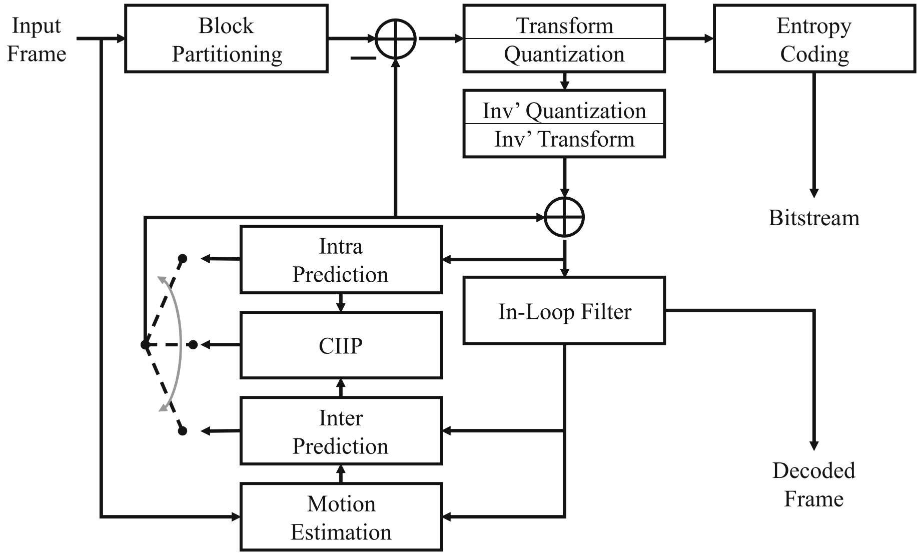


Fig. 1. Architecture of VVC/H.266

more than twice the coding efficiency of the previous standard the HEVC/H.265. Additionally, it was designed to handle Ultra-High Definition (UHD) videos ranging from 4K to 16K more efficiently and to support Virtual Reality (VR) content. As the display technology advanced, VVC/H.266 also supports High Dynamic Range (HDR), 10-/16-bit color depths as well as brightness levels of 4,000 nits and 10,000 nits. Consequently, the computational complexity significantly increased with encoding predicted to be up to 10 times higher and decoding predicted up to 2 times larger compared to HEVC/H.265.

Figure 1 illustrates the schematic overall architecture of VVC/H.266. As shown in the figure, VVC/H.266 has a block-based hybrid video coding structure that integrates various element technologies. Conceptually, video coding technology eliminates spatial, temporal, and statistical redundancies present in videos.

A video is a digital signal that quantizes continuous natural signals into discrete forms. Therefore, as continuous signals are represented by limited pixels, the performance of motion prediction for reference blocks is observed to degrade due to aliasing, rapid object movements, and other factors. Additionally, reference frames are transformed signals that have undergone quantization and inverse quantization in a block-wise manner during encoding and decoding. Although in-loop filters improve these, there are still quality differences between current frames making accurate motion prediction difficult.

To address the prediction performance degradation caused by the discontinuity of digital signals and quantization of brightness values, inter frame prediction applies low-pass filters to interpolate signals between pixels at sub-pixel levels [18]. VVC/H.266 uses the Discrete Cosine Transform-based Interpolation Filter (DCTIF) to generate sub-pixels from integer pixels. Although fractional pixels generated by interpolation filters enable more precise motion prediction, the

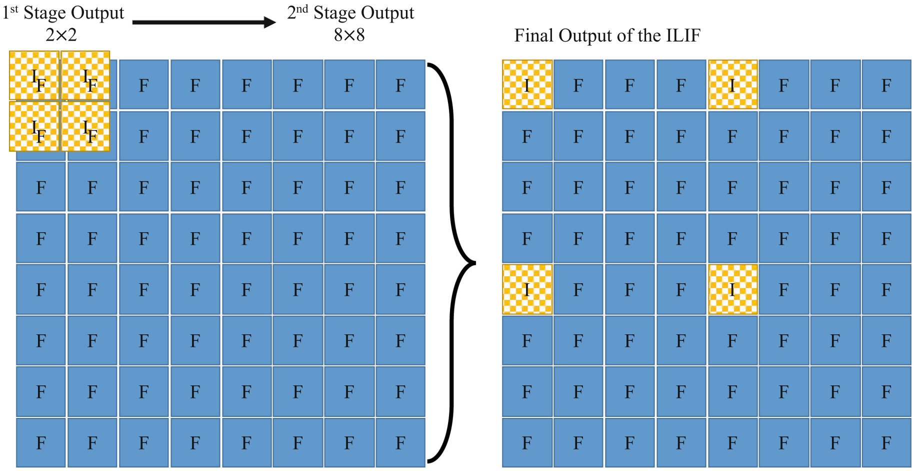


Fig. 2. Example of Integer Pixel Embedding.

input signal does not always respond ideally to handcrafted filters. Moreover, the filter coefficients are approximated to integer values for hardware optimization and high-speed computation which inherently introduces fundamental errors. These errors can become significant for certain input signals.

Recently, deep learning-based methods have shown remarkable results in image and video processing, outperforming classical methods. Video coding is also actively researching deep learning, with in-loop filters being representative examples [10]. However, several significant issues remain with the introduction of deep learning technologies for sub-pixel generation.

Firstly, there is the challenge of constructing datasets for training. NN models for image or video quality typically set unmodified original quality as the target value following general supervised learning principles. However, there are no target values for sub-pixels generated from integer pixels. To overcome this, sub-pixels generated by DCTIF are used as target values or low-pass filters are applied to the original quality images such as gaussian. Nevertheless, these approaches do not consider the sub-pixel-based inter prediction process and fail to guarantee performance in highly efficient VVC/H.266.

Secondly, the approach to sub-pixel generation as a SR problem causes issues. SR research is also an active research field in video coding, such as Reference Picture Resampling (RPR) or Video Super-Resolution (VSR) [5]. The output of SR models includes both integer and sub-pixels. In this case, the output integer pixels may not match the input integer pixels. Since sub-pixel motion estimation during the coding process is entirely dependent on input integer pixels, such discrepancies negatively impact coding efficiency.

In a real environment, the correlation between integer pixels and sub-pixels is a factor that greatly determines coding efficiency. This because of the sub-pixel motion estimation occurs after the integer pixels have been determined.

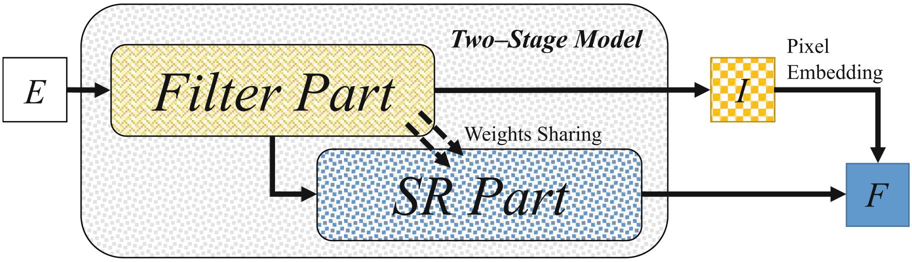


Fig. 3. Two-Stage Model using Pixel Embedding Methods.

Also, the reason why the SR model which shows a higher quality improvement surprisingly does not improve coding efficiency significantly.

In conclusion, a new approach that simultaneously consider both integer and sub-pixels during the learning and coding process is necessary unlike traditional deep learning-based sub-pixel generation research.

2 Related Work

Traditional approaches to enhance the interpolation filter have mainly focused on three aspects: enhancing fixed filters, designing adaptive filters, and developing hardware for fractional interpolation. Lakshman *et al.* have proposed a generalized interpolation framework for MCP that uses fixed-point Infinite Impulse Response (IIR) and Finite Impulse Response (FIR) filters to enhance the performance of fixed filters [9]. Wittmann *et al.* introduced the concept of a separable adaptive interpolation filter using 1-dimensional tap filters sequentially, which reduced computational cost and improved efficiency [23].

Ye *et al.* also proposed an enhanced adaptive interpolation filter that includes full pixel position filters, filter offsets, radial 12-position filters, and RD-based filter selection [26]. Lv *et al.* proposed a resolution-adaptive tap filter, selecting a 4-tap filter for high-resolution and 6-/10-tap filters for low-resolution videos [13]. Guo *et al.* proposed an efficient VLSI design by configuring the tap filters of HEVC/H.265 with an optimized parallel pipeline structure from a hardware perspective [6].

Kim and Lee suggested an 11-/12-tap Discrete Sine Transform-based Interpolation Filter (DSTIF) to emphasize high-frequency components [8]. Choi and Lee also designed a 12-tap DCTIF to improve filter response in high-frequency bands, enhancing the efficiency of VVC/H.266 [4].

As deep learning-based computer vision technology develops, attempts to apply it to interpolation filter research have been reported. Pham *et al.* have proposed a Convolutional Neural Network (CNN)-based interpolation filter for the luma and chroma components of HEVC/H.265, utilizing the sub-pixel values generated during the encoding process as the training dataset. Additionally, they applied an RDO-based interpolation filter selection method to achieve coding efficiency, adding two syntax elements for this purpose [17].

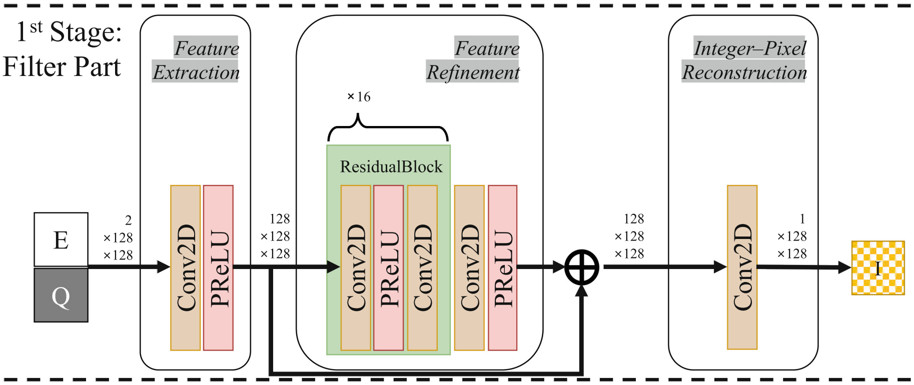


Fig. 4. Architecture of 1-st Stage Filter Model.

Similarly, Yan *et al.* used data generated by DCTIF to train a CNN model for 1/2 sub-pixel generation. They enhanced the efficiency of HEVC/H.265 by individually training 1/2 sub-pixels in vertical, horizontal, and diagonal directions [25]. In subsequent research, they proposed a CNN model for unidirectional and bidirectional motion compensation (MC), guiding the training to predict the current block to be encoded rather than generating sub-pixels from the reference block [24].

Liu *et al.* also proposed a CNN-based filter for sub-pixel interpolation in HEVC/H.265. They designed a grouped network structure for inferring sub-pixel blocks, reflecting that sub-pixel interpolation in video coding is generated at the same resolution as integer blocks. Additionally, they introduced Gaussian blurring to the target values used in training the sub-pixel generation model [11].

Murn *et al.* proposed a CNN-based interpolation filter for low-complexity inter prediction in VVC/H.266, demonstrating the potential for performance improvement [15].

Zhang *et al.* designed an interpolation filter based on the VDSR model [7], using the results of DCTIF for the dataset and introducing a constraint mask during training to maintain integer positions [27].

In this study, we prioritized maintaining the correlation and dependency between integer/fractional pixels. Likewise, recent studies on neural networks targeting video coding have attempted to utilize these kind of semantic features. Tian *et al.* suggested a framework aimed at unsupervised video semantic compression. The framework optimizes video compression by focusing on preserving semantic features rather than purely visual quality using a novel Non-Semantics Suppressed (NSS) learning strategy [20–22].

Deep learning-based interpolation filters must be designed to maintain correlation between pixels. In particular, in the case of VVC/H.266 which has very high coding efficiency, it is difficult to guarantee performance when applying a model designed for HEVC/H.265. To solve this problem, a fundamental app-

Table 1. Summary of 1-st Stage Filter Model.

Step	Layer Structure	Output Size	Parameters
Input	Luma Pixel + QP-Map	2×128×128	-
1	Conv (3×3) + PReLU	128×128×128	2,433
2∼17	Conv (3×3) + PReLU + Conv (3×3): ResBlock (×16)	128×128×128	4,722,704
18	Conv (3×3) + PReLU + Skip-Connection with Step 1	128×128×128	147,585
Output	Conv (3×3)	1×128×128	1,153
		Total Parameters	4,873,875

roach is needed that can maintain correlation during model design and learning processes.

3 In-Loop Interpolation Filter

The proposed Pixel Embedding (PE) refers to the process of directly inserting specific pixel values into specific locations in the high-resolution output image. To clarify the concept, the resolution of the input (E) and the output (F) in the SR model can be expressed as follows:

$$\begin{aligned} H_{out} &= r \times H_{in} \ , \\ W_{out} &= r \times W_{in} \ , \end{aligned} \tag{1}$$

Here, (H_{in}, W_{in}) and (H_{out}, W_{out}) are the resolutions of the input image and output image, respectively. r is the scaling factor of the SR model. Therefore, the equation for embedding E into the integer pixel positions of F can be expressed as follows:

$$F(ri, rj) = E(i, j) \quad \text{for} \quad 0 \leq i < H_{in} \ , \ 0 \leq j < W_{in} \ . \tag{2}$$

All pixel values $F(i, j)$ in the output are used to calculate the loss function with the target values during the training process. Since the pixel value $F(ri, rj)$ is always equal to the input pixel $E(i, j)$ of the model, the correlation between the integer pixels and the sub-pixels is maintained even as training progresses.

However, the limitation of this approach is that since the value $F(ri, rj)$ does not change, it is difficult to expect an overall improvement in high-resolution quality due to training. Originally, SR models aim for both quality improvement and up-scaling simultaneously, so this kind of constraint needs to be improved.

The quality of all pixels $F(i, j)$ in the output image is constrained by the quality of the fixed pixels $F(ri, rj)$. Therefore, if the quality of $F(ri, rj)$ is improved compared to the input pixel $E(i, j)$, the quality of all pixels $F(i, j)$ can also be improved. From this perspective, we can introduce a filter model $\mathrm{V}(\cdot)$ that generates a high-quality output (I) with the same resolution as the input image (E). So, we can modify the equation as follows:

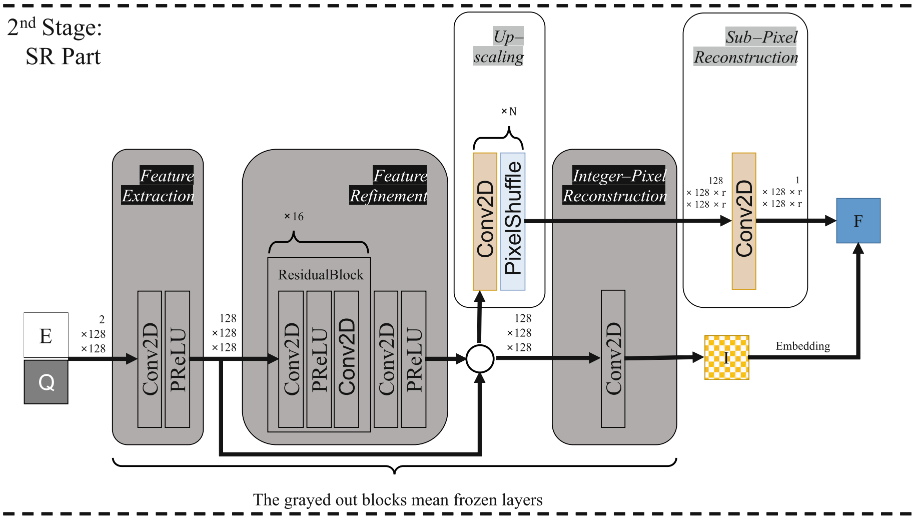


Fig. 5. 2-nd Stage Model.

$$
\begin{aligned}
I(i,j) &= \mathrm{V}(E)(i,j)\ ,\\
F(ri,rj) &= I(i,j) \quad \text{for} \quad 0 \le i < H_{in}\ ,\ 0 \le j < W_{in}\ .
\end{aligned}
\tag{3}
$$

By changing the target of pixel embedding to the output of the filter model $\mathrm{V}(\cdot)$, the quality of the output pixels $F(i,j)$ from the SR model is also improved compared to the previous results. Thus, this approach achieves improved quality of integer pixels, generation of high-quality sub-pixels, and maintains the dependency between integer and sub-pixels through pixel embedding.

Figure 2 details this process. **I** represents the integer pixel samples output from the filter model, exemplified as having a size of 2×2. **F** represents the sub-pixel samples output from the SR model, exemplified as having a size of 8×8, which is four times the input resolution. The final output maintains the correlation between integer and sub-pixels by embedding **I** into the integer pixel positions of **F**. Through the pixel embedding process, the quality of the optimized integer pixel samples from filter model is preserved Also, high-quality sub-pixel samples with well-preserved correlation to the integer pixels are generated.

However, introducing the filter model $\mathrm{V}(\cdot)$ and the SR model separately can lead to additional issues. To reduce the complexity of the neural network model, we did not consider an ensemble approach of separate models. Therefore, we adopted a two-stage training strategy that divides the network module to achieve filtering and SR with a single model simultaneously.

Figure 3 illustrates the Pixel Embedding process through the proposed two-stage model. We have employed a two-stage learning method that separates the filter part and the SR part within a single network and optimized them sepa-

Table 2. Summary of 2-nd Stage SR Model.

Step	Layer Structure	Output Size	Parameters
Input	Luma Pixel + QP-Map	2×128×128	-
1	Conv (3×3) + PReLU	128×128×128	2,433
2∼17	Conv (3×3) + PReLU + Conv (3×3): ResBlock (×16)	128×128×128	4,722,704
18	Conv (3×3) + PReLU + Skip-Connection with Step 1	128×128×128	147,585
19	Conv (3×3)	1×128×128	1,153
Filter Part Parameters			4,873,875
20	Conv (3×3)	512×128×128	590,336
21	PixelShuffle (2×)	128×256×256	-
22	Conv (3×3)	512×256×256	590,336
23	PixelShuffle (2×)	128×512×512	-
24	Conv (3×3)	512×512×512	590,336
25	PixelShuffle (2×)	128×1,024×1,024	-
26	Conv (3×3)	512×1,024×1,024	590,336
27	PixelShuffle (2×)	128×2,048×2,048	-
28	Conv (3×3)	1×2,048×2,048	1,153
Output	Pixel Embedding: $F(ri, rj) = I(i, j)$	1×2,048×2,048	-
SR Part Parameters			2,362,497
Total Parameters			7,236,372

rately. The two parts are optimized in separate stages, and the trained weights of the filter part are fully shared in the SR part. This approach achieves the same goal not using separate filter and SR models but with a single network.

As shown in Fig. 4, 1-st stage model functions as a typical filtering model that performs an E2E mapping of low-quality input pixels to high-quality output pixels. Thus, the goal of the first stage is to fine-tune the model to improve the quality of the input image, excluding the up-scaling part. Table 1 summarizes the configuration of the 1-st stage model and the dimensions and number of parameters of the feature map output from each layer. The 1-st stage model is a filter model with 4,873,875 parameters.

As shown in Fig. 5, this 2-nd stage model tunes the up-scaling part to generate sub-pixel samples. All parameters tuned in the 1-st stage (*Feature Extraction, Feature Refinement, Integer-Pixel Reconstruction*) are transferred and shared. Also, these are frozen and excluded from the weight update process in the 2-nd stage learning. This means that the integer pixel output from the 1-st stage does not change at this point. However, since the integer pixel part included in the output of the this stage changes, the output values of the 1-st stage are embedded into the output values of the 2-nd stage to correct this in the final output.

Discontinuity between integer pixels and subpixels caused by embedding naturally disappear during the process of backpropagating the loss function. This is because the loss function is calculated over the entire plane containing

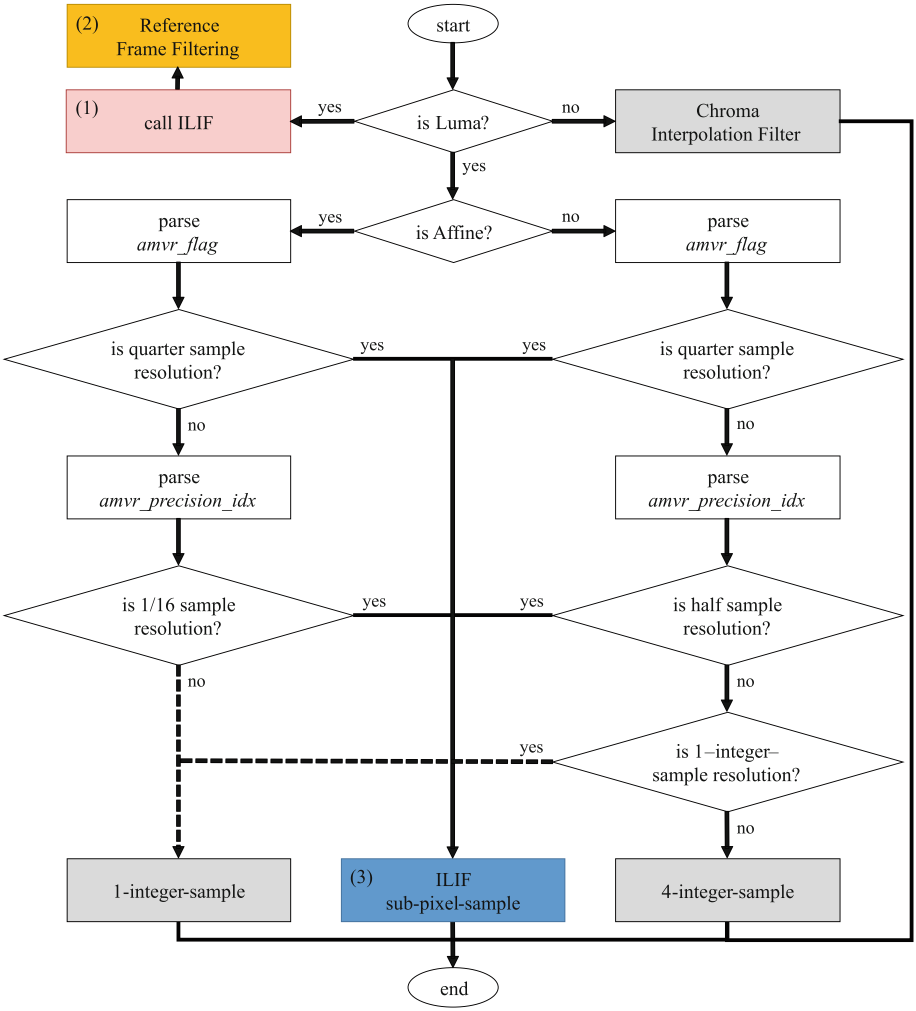

(2) Reference Frame Filtering
start
(1) call ILIF
yes
is Luma?
no
Chroma Interpolation Filter
yes
parse *amvr_flag*
yes
is Affine?
no
parse *amvr_flag*
is quarter sample resolution?
yes
yes
is quarter sample resolution?
no
no
parse *amvr_precision_idx*
parse *amvr_precision_idx*
is 1/16 sample resolution?
yes
yes
is half sample resolution?
no
no
yes
is 1–integer–sample resolution?
no
1-integer-sample
(3) ILIF sub-pixel-sample
4-integer-sample
end

Fig. 6. Flow Chart of the proposed Inter Prediction with ILIF.

the embedded integer pixels. Table 2 summarizes the configuration of the 2-nd stage model and the dimensions and number of parameters of the feature map output from each layer. The 2-nd stage model is an SR model with 7,236,372 parameters. Among them, excluding the parameters shared and freezed from the 1-st stage model, the number of parameters is 2,362,497.

The proposed ILIF simultaneously performs the roles of an in-loop filter and an interpolation filter in the inter prediction process targeting VVC/H.266. Thus, there are 2 modes for integrating ILIF into the VVC/H.266 inter prediction process as follows:

- MODE 1: Sampling Fractional Pixels of ILIF for DCTIF Replacement:

  - Use only the fractional-pixel output F from ILIF to replace the DCTIF
  - This method enhances the motion compensation accuracy by providing high-quality fractional pixels
- MODE 2: Combined Integer Pixel Filtering and Fractional Pixel Sampling:
  - Use both the integer pixel output I to filter the reference frame and the fractional-pixel output F to replace DCTIF
  - This combined approach leverages the strengths of both integer pixel filtering and fractional-pixel generation for optimal performance

Figure 6 illustrates the inter prediction process within VVC/H.266 integrated with ILIF. The components numbered as (1), (2), (3) in the figure represent the additional logic introduced with ILIF integration. The ILIF model is trained considering only the luma components. During the inter prediction stage, the ILIF model is called to generate fractional samples at 16 times the size when the reference sample is a luma component as shown in component (1). The generated fractional samples include enhanced integer pixel samples.

MODE 1 corresponds to using the fractional samples generated by ILIF as shown in component (3) for 1/4 and 1/16 level motion compensation (MC) in AFFINE AMVP mode or 1/2 and 1/4 level MC in Normal AMVP mode. In MODE 2, both components (2) and (3) correspond to the active state. Therefore, the overall efficiency may be further improved since motion prediction becomes more accurate from the integer pixel unit.

The proposed ILIF-based inter prediction technique is designed with the goal of achieving high performance in MODE 3, even if the performance or gains in MODE 2 are lower or minimal. This strategy is based on the fundamental design of DCTIF which maintains the correlation between integer and fractional pixels Also, the interdependence of the inter prediction process were considered. Ultimately, proposed ILIF can contribute to improving the efficiency of inter prediction in VVC/H.266 through the integration of in-loop filter and interpolation filter.

4 Experimental Results

We used two main datasets for training ILIF. First, we used 22 sequences from Class A to E of the Common Test Condition (CTC) for VVC/H.266 [1]. Second, the sequence dataset from Bristol University (BVI-DVC) including four different resolutions, ranging from 270p to 2160p, with 200 sequences for each resolution, totaling 800 sequences [14]. The primary purpose of using the BVI-DVC dataset is for the 2-nd stage of ILIF training. For the purpose, the original sequences of BVI-DVC were first downsampled to 1/4 of their size before being encoded.

The CTC sequences and the 1/4 downsampled BVI-DVC sequences were encoded using the VVC Test Model version 11.0 (VTM-11.0) which is the reference software for VVC/H.266 [3]. For encoding, the Random Access (RA) configuration file (*encoder_randomaccess_vtm.cfg*) was used.

ILIF was trained as a single integrated model not by QP but using normalized QP-Map [19] for the input samples. To train the 256 times output model using datasets where most sequences have resolutions smaller than 4K, we utilized the DCTIF coefficients of VVC/H.266. We normalized the integer-scaled 8-tap luma DCTIF coefficients and then constructed 9-tap filter coefficients to ensure symmetry. During training, we applied horizontal and vertical filtering to the target values t to generate new target values $\hat{t}$ and calculated the loss between these and the outputs y of the SR Part. This approach has the advantage of fully utilizing the CTC and BVI-DVC sequences without the need for a downsampling process. Additionally, it offers benefits in both original quality and fractional sample generation by applying DCTIF to the original target values.

The proposed ILIF is designed with the goal of simultaneously improving the visual quality of integer pixels and generating ultra-high-resolution sub-pixels. For the training of the 2-nd stage model, we generated the upscaled target values from the uncompressed original frames using the normalized DCTIF ($\mathcal{D}$).

Given an original target pixel data t, we apply a convolution process to upscale the data both horizontally and vertically by a factor of $r = 16$. The process involves the following steps:

$$\begin{aligned} t_h(i, rj + c) &= \sum_{k=-4}^{4} t(i, j + k) \cdot \mathcal{D}(c, k + 4) \ , \\ \hat{t}(ri + c, rj + c) &= \sum_{k=-4}^{4} t_h(i + k, rj + c) \cdot \mathcal{D}(c, k + 4) \ . \end{aligned} \tag{4}$$

Here, t_h is the result after horizontal convolution, of size $(H \times (W \times r))$ and $\hat{t}$ is the final upscaled result, of size $(H \times r) \times \ (\mathrm{W} \times r)$. $\mathcal{D}$ is a 16×9 matrix containing the nomalized filter coefficients. c is the index value for 16 tap filters assigned by sub-pixel generation positions. The 16 sets of normalized 9-tap luma DCTIF filter coefficients $\mathcal{D}$ as follows:

$$\mathcal{D} = \begin{pmatrix}
0.0000 & 0.0000 & 0.0000 & 0.0000 & 1.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 \\
0.0000 & 0.0000 & 0.0156 & -0.0469 & 0.9844 & 0.0625 & -0.0313 & 0.0156 & 0.0000 \\
0.0000 & -0.0156 & 0.0313 & -0.0781 & 0.9688 & 0.1250 & -0.0469 & 0.0156 & 0.0000 \\
0.0000 & -0.0156 & 0.0469 & -0.1250 & 0.9375 & 0.2031 & -0.0625 & 0.0156 & 0.0000 \\
0.0000 & -0.0156 & 0.0625 & -0.1563 & 0.9063 & 0.2656 & -0.0781 & 0.0156 & 0.0000 \\
0.0000 & -0.0156 & 0.0625 & -0.1719 & 0.8125 & 0.4063 & -0.1250 & 0.0469 & -0.0156 \\
0.0000 & -0.0156 & 0.0469 & -0.1406 & 0.7344 & 0.4844 & -0.1563 & 0.0625 & -0.0156 \\
0.0000 & -0.0156 & 0.0625 & -0.1719 & 0.7031 & 0.5313 & -0.1563 & 0.0625 & -0.0156 \\
0.0000 & -0.0156 & 0.0625 & -0.1719 & 0.6250 & 0.6250 & -0.1719 & 0.0625 & -0.0156 \\
0.0000 & -0.0156 & 0.0625 & -0.1563 & 0.5313 & 0.7031 & -0.1719 & 0.0625 & -0.0156 \\
0.0000 & -0.0156 & 0.0625 & -0.1563 & 0.4844 & 0.7344 & -0.1406 & 0.0469 & -0.0156 \\
0.0000 & -0.0156 & 0.0469 & -0.1250 & 0.4063 & 0.8125 & -0.1719 & 0.0625 & -0.0156 \\
0.0000 & 0.0000 & 0.0156 & -0.0781 & 0.2656 & 0.9063 & -0.1563 & 0.0625 & -0.0156 \\
0.0000 & 0.0000 & 0.0156 & -0.0625 & 0.2031 & 0.9375 & -0.1250 & 0.0469 & -0.0156 \\
0.0000 & 0.0000 & 0.0156 & -0.0469 & 0.1250 & 0.9688 & -0.0781 & 0.0313 & -0.0156 \\
0.0000 & 0.0000 & 0.0156 & -0.0313 & 0.0625 & 0.9844 & -0.0469 & 0.0156 & 0.0000
\end{pmatrix} \tag{5}$$

We used Mean Absolute Difference (MAD) for the cost function. Therefore, the loss function was defined as follows:

$$\mathcal{L}_{16} = \frac{1}{N}\sum_{i=1}^{N} w_{qp} \cdot |y_i - \hat{t}_i| \ . \tag{6}$$

Here, w_{qp} is a weight value according to the base QP of the dataset and was introduced to prevent the model from overfitting to data of a specific QP. The weight values for each of the 5 base QPs (22, 27, 32, 37, and 42) were set to 1.7, 1.5, 1.3, 1.1, and 1.0, respectively.

PyTorch was used as the framework for implementing the proposed ILIF [16]. We used Adaptive Moment Estimation with Weight Decay (ADAMW) optimization algorithm for training [12]. The training utilized a multi-GPU environment with 4 GPUs.

To verify the performance of ILIF, we modified the VVC reference software (VTM-11.0) and integrated them using LibTorch, the C++ API of PyTorch. We encoded the Class A$\sim$E sequences of the CTC using the RA Main 10 configuration. All results were presented as BD-rate performance for 50 frames per sequence between proposed ILIF the VVC/H.266 (VTM-11.0) anchor. Both the training of the ILIF model and the integration with the reference software considered only the luma (Y) component. Therefore, only the results for Y are significant in the experimental results.

The experimental results for the integration of ILIF and VVC covered whole AMVR resolution are summarized in Tables 3, and 4, respectively. These results are implemented to support up to 1/16-luma-sample resolution which is used in the AFFINE AMVP mode.

MODE 1 presents the results when the proposed ILIF is applied only for sub-pixel generation. The experimental results show the gains of -1.42% for All-QP, -1.54% for High-QP, and 1.24% for Low-QP. It can be seen that the performance improvement due to AFFINE mode and 1/16 ultra-high resolution sub-pixel samples is significant.

MODE 2 involves applying the proposed ILIF for both integer pixel improvement and sub-pixel generation. The experimental results show the gains of -3.92% for All-QP, -4.01% for High-QP, and -3.13% for Low-QP, respectively. The performance improvement in MODE 2 can be attributed to maintaining the dependency between integer and sub-pixels.

Table 3. BD-Rate Comparison of the ILIF: **MODE 1** (1/16-luma-sample).

Random Access Main 10 (50 frames, 1/16-luma-sample)										
Class	Sequence	All-QP			High-QP			Low-QP		
		{22, 27, 32, 37, 42}			{27, 32, 37, 42}			{22, 27, 32, 37}		
		Y	U	V	Y	U	V	Y	U	V
A1	Tango2	−0.70%	2.50%	3.40%	−0.80%	2.30%	3.50%	−0.60%	2.00%	3.30%
	FoodMarket4	−1.10%	2.70%	3.10%	−1.20%	2.80%	3.20%	−1.00%	2.60%	3.00%
	Campfire	0.50%	1.30%	1.40%	0.40%	1.40%	1.50%	0.60%	1.20%	1.30%
A2	CatRobot	−4.40%	3.10%	3.00%	−4.50%	3.20%	3.10%	−4.30%	3.00%	2.90%
	DaylightRoad2	−5.00%	3.30%	3.10%	−5.10%	3.40%	3.20%	−4.90%	3.20%	3.00%
	ParkRunning3	−0.10%	1.50%	1.60%	−0.20%	1.60%	1.70%	−0.10%	1.40%	1.50%
B	MarketPlace	−0.80%	3.40%	3.50%	−0.90%	3.30%	3.60%	−0.70%	3.20%	3.40%
	RitualDance	−0.90%	1.80%	2.20%	−1.00%	1.70%	2.30%	−0.80%	1.90%	2.10%
	Cactus	−3.30%	3.50%	3.60%	−3.40%	3.40%	3.70%	−3.20%	3.60%	3.50%
	BasketballDrive	−0.70%	3.70%	3.40%	−0.80%	3.60%	3.50%	−0.60%	3.80%	3.30%
	BQTerrace	−9.80%	2.60%	2.70%	−9.90%	2.50%	2.80%	−9.70%	2.70%	2.60%
C	BasketballDrill	−1.10%	3.22%	2.35%	−1.24%	3.93%	2.68%	−0.96%	2.64%	1.94%
	BQMall	−0.11%	2.03%	2.74%	−0.28%	2.28%	3.46%	0.27%	1.76%	2.11%
	PartyScene	−0.12%	1.61%	1.75%	−0.36%	2.15%	1.98%	0.14%	1.25%	1.53%
	RaceHorses	1.62%	2.04%	2.90%	1.99%	2.17%	3.78%	1.39%	1.75%	2.10%
D	BasketballPass	2.30%	4.01%	2.52%	2.57%	4.95%	2.72%	2.01%	3.54%	2.22%
	BQSquare	−1.25%	3.28%	3.43%	−1.94%	3.36%	3.25%	−0.68%	3.22%	3.33%
	BlowingBubbles	1.29%	2.14%	2.65%	1.23%	2.51%	2.90%	1.30%	2.02%	2.25%
	RaceHorses	2.70%	2.25%	2.25%	2.87%	2.28%	2.09%	2.50%	1.63%	1.21%
E	FourPeople	−4.07%	2.08%	1.88%	−4.51%	2.34%	1.93%	−2.94%	1.94%	1.62%
	Johnny	−4.50%	2.49%	2.81%	−4.67%	2.84%	3.14%	−4.13%	2.13%	2.41%
	KristenAndSara	−1.69%	3.01%	2.53%	−2.24%	3.48%	2.71%	−0.77%	2.65%	2.24%
Class A1		−0.77%	2.17%	2.63%	−0.87%	2.30%	2.93%	−0.73%	2.00%	2.57%
Class A2		−4.28%	2.61%	2.92%	−4.40%	2.68%	3.04%	−4.17%	2.39%	2.76%
Class B		−3.32%	2.95%	3.34%	−3.49%	2.96%	3.49%	−3.14%	2.98%	3.32%
Class C		0.07%	2.22%	2.44%	0.03%	2.63%	2.97%	0.21%	1.85%	1.92%
Class D		1.26%	2.92%	2.72%	1.18%	3.27%	2.74%	1.28%	2.60%	2.25%
Class E		−3.42%	2.53%	2.40%	−3.81%	2.89%	2.59%	−2.62%	2.24%	2.09%
Overall		−1.42%	2.62%	2.67%	−1.54%	2.80%	2.85%	−1.24%	2.42%	2.40%

Table 4. BD-Rate Comparison of the ILIF: **MODE 2** (1/16-luma-sample).

Random Access Main 10 (50 frames, 1/16-luma-sample)										
Class	Sequence	All-QP			High-QP			Low-QP		
		{22, 27, 32, 37, 42}			{27, 32, 37, 42}			{22, 27, 32, 37}		
		Y	U	V	Y	U	V	Y	U	V
A1	Tango2	−1.21%	1.49%	2.60%	−1.22%	1.38%	2.82%	−1.40%	1.10%	2.17%
	FoodMarket4	−2.18%	1.65%	2.02%	−2.21%	2.01%	2.49%	−2.21%	1.11%	1.62%
	Campfire	−0.48%	0.23%	0.51%	−0.76%	0.42%	0.78%	−0.28%	0.14%	0.36%
A2	CatRobot	−5.49%	2.13%	1.95%	−5.46%	2.36%	2.16%	−5.31%	1.41%	1.42%
	DaylightRoad2	−6.04%	2.19%	1.98%	−6.71%	2.29%	2.25%	−5.89%	1.58%	1.67%
	ParkRunning3	−1.14%	0.40%	0.65%	−1.34%	0.46%	0.81%	−1.05%	0.34%	0.48%
B	MarketPlace	−1.33%	2.36%	2.45%	−1.28%	2.36%	2.71%	−1.52%	2.12%	1.76%
	RitualDance	−1.46%	0.75%	1.23%	−1.66%	1.04%	1.24%	−1.33%	0.44%	0.82%
	Cactus	−4.34%	2.54%	2.48%	−4.86%	2.78%	2.58%	−3.50%	2.15%	2.20%
	BasketballDrive	−1.74%	2.65%	2.37%	−2.03%	2.62%	2.13%	−1.43%	2.16%	2.03%
	BQTerrace	−10.76%	1.57%	1.68%	−14.72%	1.56%	1.73%	−8.52%	1.73%	1.68%
C	BasketballDrill	−4.59%	0.29%	0.20%	−5.32%	0.26%	−0.11%	−4.00%	0.31%	0.19%
	BQMall	−4.04%	0.15%	0.92%	−4.99%	0.06%	1.02%	−2.99%	0.10%	0.60%
	PartyScene	−3.18%	−0.14%	0.38%	−4.48%	−0.37%	0.26%	−2.12%	−0.20%	0.35%
	RaceHorses	−0.03%	0.43%	0.28%	0.05%	0.25%	0.66%	−0.09%	0.33%	−0.06%
D	BasketballPass	−0.44%	1.30%	0.97%	−1.36%	1.12%	0.74%	0.23%	1.47%	1.29%
	BQSquare	−7.04%	0.16%	0.31%	−8.20%	0.08%	−0.02%	−6.11%	0.17%	0.37%
	BlowingBubbles	−1.29%	−0.14%	−0.14%	−1.77%	−0.25%	−0.03%	−0.93%	0.24%	−0.31%
	RaceHorses	0.21%	−1.02%	−0.55%	0.18%	−1.33%	−0.90%	0.06%	−0.41%	−0.62%
E	FourPeople	−7.96%	0.71%	0.71%	−8.95%	0.57%	0.55%	−6.03%	0.83%	0.61%
	Johnny	−9.22%	0.47%	1.09%	−9.99%	0.51%	0.79%	−8.01%	0.31%	1.21%
	KristenAndSara	−7.28%	1.01%	0.78%	−8.45%	1.01%	0.69%	−5.32%	0.97%	0.78%
Class A1		−1.29%	1.12%	1.71%	−1.40%	1.27%	2.03%	−1.29%	0.78%	1.38%
Class A2		−4.22%	1.57%	1.53%	−4.50%	1.70%	1.74%	−4.08%	1.11%	1.19%
Class B		−3.92%	1.97%	2.04%	−4.91%	2.07%	2.08%	−3.26%	1.72%	1.70%
Class C		−2.96%	0.18%	0.45%	−3.68%	0.05%	0.46%	−2.30%	0.14%	0.27%
Class D		−2.14%	0.07%	0.15%	−2.79%	−0.10%	−0.05%	−1.69%	0.37%	0.18%
Class E		−8.15%	0.73%	0.86%	−9.13%	0.70%	0.68%	−6.45%	0.70%	0.87%
Overall		−3.92%	0.97%	1.22%	−4.01%	0.90%	1.28%	−3.13%	0.88%	0.96%

5 Conclusion

The ultimate goal of this dissertation is to propose a deep learning-based SR model to improve the efficiency of inter-coding within video coding standards and apply it simultaneously for integer pixel enhancement and sub-pixel generation.

The proposed ILIF was an SR model with a fully convolutional neural network structure containing 16 Residual Blocks and 4 Pixel Shuffling Blocks. In addition, this model divided the network into a filtering part and an SR part to achieve the aforementioned dual objectives. The output of the filtering part and the output of the SR part were combined into one output but, unlike common SR models the Pixel Embedding techniques were utilized to maintain correlation. Pixel Embedding was a method of embedding the integer pixel output of the filtering part directly into the sub-pixel output of the SR part. This induced the sub-pixels to correlate to the integer pixels.

ILIF totally replaced DCTIF in the inter prediction technology of VVC/H.266 and was used as a high-performance interpolation filter. Depending on the utilization of integer and sub-pixel samples generated from ILIF, it was categorized 2 integration methods with VVC/H.266 as named MODE 1 and 2.

The experimental results showed improvement in performance was observed when ILIF was applied up to 1/16-luma-sample resolution. The results showed the gains of −1.42% for All-QP, −1.54% for High-QP, and −1.24% for Low-QP in MODE 1 which applies ILIF only for sub-pixel generation However, significant BD-rate gains were observed as −3.92% for All-QP, −4.01% for High-QP, and −3.13% for Low-QP in MODE 2 which integrates both integer pixel filtering and sub-pixel generation.

The comprehensive experimental results demonstrated that incorporating both integer pixels and sub-pixels in the learning model could enhance the performance of inter prediction techniques.

References

1. Bossen, F., Boyce, J., Suehring, K., Li, X., Seregin, V.: VTM common test conditions and software reference configurations for SDR video. document JVET-T2010, Teleconference, October 2020
2. Bross, B., et al.: Overview of the Versatile Video Coding (VVC) standard and its applications. IEEE Trans. Circuits Syst. Video Technol. **31**(10), 3736–3764 (2021)
3. Chen, J., Ye, Y., Kim, S.H.: Algorithm description for Versatile Video Coding and Test Model 11 (VTM 11). document JVET-T2002, Teleconference, October 2020
4. Choi, M.K., Lee, Y.L.: Adaptive interpolation filter using correlation for inter prediction. IEEE Access **11**, 131017–131023 (2023)
5. Choi, Y.J., Kim, B.G.: Hirn: hierarchical recurrent neural network for video super-resolution (vsr) using two-stage feature evolution. Appl. Soft Comput. **143**, 110422 (2023)
6. Guo, Z., Zhou, D., Goto, S.: An optimized mc interpolation architecture for hevc. In: 2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 1117–1120 (2012)
7. Kim, J., Lee, J.K., Lee, K.M.: Accurate image super-resolution using very deep convolutional networks. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1646–1654 (2015)
8. Kim, M., Lee, Y.L.: Discrete sine transform-based interpolation filter for video compression. Symmetry **9**, 257 (2017)
9. Lakshman, H., Schwarz, H., Wiegand, T.: Generalized interpolation-based fractional sample motion compensation. IEEE Trans. Circuits Syst. Video Technol. **23**, 455–466 (2013)
10. Lee, Y.W., Kim, J.H., Choi, Y.J., Kim, B.G.: Cnn-based approach for visual quality improvement on hevc. In: 2018 IEEE International Conference on Consumer Electronics (ICCE), pp. 1–3 (2018)

11. Liu, J., Xia, S., Yang, W., Li, M., Liu, D.: One-for-all: grouped variation network-based fractional interpolation in video coding. IEEE Trans. Image Process. **28**, 2140–2151 (2019)
12. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. In: International Conference on Learning Representations (2017)
13. Lv, H., Wang, R., Li, Y., Zhu, C., Jia, H., Xie, X., Gao, W.: A resolution-adaptive interpolation filter for video codec. In: 2014 IEEE International Symposium on Circuits and Systems (ISCAS), pp. 542–545 (2014)
14. Ma, D., Zhang, F., Bull, D.R.: BVI-DVC: a training database for deep video compression. IEEE Trans. Multimedia **24**, 3847–3858 (2022)
15. Murn, L., Blasi, S.G., Smeaton, A.F., Mrak, M.: Improved cnn-based learning of interpolation filters for low-complexity inter prediction in video coding. IEEE Open J. Signal Process. **2**, 453–465 (2021)
16. Paszke, A., et al.: PyTorch: an imperative style, high-performance deep learning library. In: Advances in Neural Information Processing Systems, vol. 32. Curran Associates, Inc. (2019)
17. Pham, C.D.K., Zhou, J.: Deep learning-based luma and chroma fractional interpolation in video coding. IEEE Access **7**, 112535–112543 (2019)
18. Richardson, I.E.G.: The mpeg-4 and h.264 standards (2004)
19. Song, X., Yao, J., Zhou, L., Wang, L., Wu, X., Xie, D., Pu, S.: A Practical Convolutional Neural Network as Loop Filter for Intra Frame. In: 2018 25th IEEE International Conference on Image Processing (ICIP), pp. 1133–1137 (2018)
20. Tian, Y., Lu, G., Yan, Y., Zhai, G., Chen, L., Gao, Z.: A coding framework and benchmark towards low-bitrate video understanding. IEEE Trans. Pattern Anal. Mach. Intell. **46**, 5852–5872 (2024)
21. Tian, Y., Lu, G., Zhai, G., Gao, Z.: Non-semantics suppressed mask learning for unsupervised video semantic compression. 2023 IEEE/CVF International Conference on Computer Vision (ICCV), pp. 13564–13576 (2023)
22. Tian, Y., Yan, Y., Zhai, G., Chen, L., Gao, Z.: Clsa: a contrastive learning framework with selective aggregation for video rescaling. IEEE Trans. Image Process. **32**, 1300–1314 (2023)
23. Wittmann, S., Wedi, T.: Separable adaptive interpolation filter for video coding. In: 2008 15th IEEE International Conference on Image Processing, pp. 2500–2503 (2008)
24. Yan, N., Liu, D., Li, H., Li, B., Li, L., Wu, F.: Convolutional neural network-based fractional-pixel motion compensation. IEEE Trans. Circuits Syst. Video Technol. **29**, 840–853 (2019)
25. Yan, N., Liu, D., Li, H., Wu, F.: A convolutional neural network approach for half-pel interpolation in video coding. In: 2017 IEEE International Symposium on Circuits and Systems (ISCAS), pp. 1–4 (2017)
26. Ye, Y., Motta, G., Karczewicz, M.: Enhanced adaptive interpolation filters for video coding. In: 2010 Data Compression Conference, pp. 435–444 (2010)
27. Zhang, H., Song, L., Luo, Z., Yang, X.: Learning a convolutional neural network for fractional interpolation in hevc inter coding. In: 2017 IEEE Visual Communications and Image Processing (VCIP), pp. 1–4 (2017)

Scene Text Image Super-Resolution with CLIP Prior Guidance

Yogesh Surapaneni(✉) and Chakravarthy Bhagvati

School of Computer & Information Sciences, University of Hyderabad, Hyderabad, India
surapaneniyogesh11@gmail.com, chakravarthybhagvati@uohyd.ac.in
https://scis.uohyd.ac.in/

Abstract. Scene Text Image Super-Resolution (STISR) plays a crucial role in enhancing text readability within natural scenes, impacting OCR systems, visual question answering, and image retrieval. Existing STISR methods often fall short, either neglecting textual information entirely or utilizing it ineffectively. We attempt to bridge this gap with a novel two-fold approach. Firstly, we use CLIP (Contrastive Language-Image Pre-Training), a powerful model that can map both images and text into a shared latent space, enabling it to assess image-text alignment. We harness CLIP's ability to understand the semantic relationship between images and text. By incorporating CLIP-generated image representations that capture these inherent textual features, we effectively guide the super-resolution process, leading to more accurate reconstructions. Secondly, we propose a novel TrOCR (Transformer-based OCR) loss function to supervise the super-resolution process from a text-centric perspective. Our loss function enforces consistency between the super-resolved output and the high-resolution ground truth image in terms of their text content. Experiments conducted on the benchmark TextZoom dataset demonstrate that our approach not only improves visual quality but also boosts text recognition accuracy.

Keywords: Scene Text Image Super-Resolution · CLIP Embeddings · StyleGAN2 · TrOCR Loss · Deep Learning

1 Introduction

Digital images are a cornerstone of modern information transmission, yet limitations in image acquisition systems can lead to images with insufficient detail or resolution. Super-Resolution (SR) techniques address this by reconstructing high-resolution (HR) images from low-resolution (LR) inputs.

Early SR methods relied on interpolation techniques, but these often resulted in artifacts such as blurring and aliasing due to their inherent inability to introduce new information corresponding to the finer details that appear at high resolution. More recent advancements incorporate learning-based approaches, utilizing large datasets of paired LR and HR images to train deep neural networks. These methods, such as SRCNN [6], VDSR [13], and SRGAN [15], have shown significant improvements by intelligently filling in plausible details and maintaining the natural characteristics of images.

A. Antonacopoulos et al. (Eds.): ICPR 2024, LNCS 15332, pp. 17–32, 2025.
https://doi.org/10.1007/978-3-031-78125-4_2

However, while significant progress has been made in enhancing natural images, applying SR to scene text images presents unique challenges. Scene Text Image Super-Resolution (STISR) focuses on improving the resolution and legibility of text within complex backgrounds, varying illuminations, and diverse fonts and styles. Early methods directly used generic SR approaches and ignored these text characteristics in scene text images. Then STISR methods like TSRN (Text Super Resolution Network) [28] and TPGSR (Text Prior Guided Super-Resolution) [23] started incorporating text-specific features and loss functions, but they still struggle to extract and harness the full potential of the textual features.

Our work aims to bridge this gap by proposing a novel STISR approach that addresses the unique challenges of scene text images. Our key contributions are as follows:

1. We explore the use of CLIP for STISR, highlighting its strengths in text feature extraction through fine-tuning with literal text pairings.
2. We introduce the TrOCR loss function, designed to improve both visual quality and text recognition accuracy. The effectiveness of TrOCR is demonstrated by its ability to elevate these metrics when applied to existing STISR methods.
3. We propose a comprehensive new method by modifying StyleGAN2, integrating CLIP image embeddings, and introducing our novel TrOCR loss, enhancing scene text image resolution and readability.

2 Related Work

2.1 Single Image Super Resolution (SISR)

SISR techniques aim to reconstruct HR images from LR inputs. Early methods like SRCNN [6] introduced the potential of deep learning with a three-layer CNN. VDSR [13] expanded on this with deeper networks, while SRGAN [15] utilized generative adversarial networks (GANs) and perceptual loss for more realistic images. Subsequent models, such as EDSR [18], RDN [31], LapSRN [14], and RCAN [30], optimized network design and training for better performance and efficiency. The advent of transformers further advanced SISR with models like IPT [3] and SwinIR [17], showcasing state-of-the-art performance. Other approaches like AND [21] and FuncNet [20] have further improved SISR by addressing degradation robustness and parametric restoration, respectively.

2.2 Scene Text Image Super-Resolution (STISR)

STISR enhances text resolution in natural scenes, addressing challenges like blurred characters and distorted shapes. Early research, exemplified by TSRN [28], highlighted limitations in generic SISR for text data. TSRN employs sequential residual blocks and a boundary-aware loss function to enhance character flow and sharpness. TPGSR [23] integrates text recognition models to generate “character probability sequences”, improving reconstruction accuracy.

Adversarial learning methods like TSRGAN [7] maintain text spatial structure through the Sinkhorn distance and enhance visual realism with triplet attention. STT [4] uses Transformers for accurate character reconstruction despite distortions. Text Gestalt [5] prioritizes stroke clarity with a Stroke-Focused Module (SFM), while TATT [22] employs global attention mechanisms for spatial coherence. C3-STISR [32] integrates visual, textual, and linguistic features for enhanced reconstructions, and DPMN [34] refines text and graphic recognition priors to modulate super-resolution for improved visual and textual clarity. Additionally, text-conditional diffusion models [24] have been proposed, utilizing their powerful text-to-image synthesis capabilities to significantly surpass existing STISR methods, particularly in producing superior quality super resolution text images.

2.3 Scene Text Recognition (STR)

Scene Text Recognition (STR) deals with deciphering text in natural images, facing challenges like variable fonts, orientations, and occlusions. Unlike Optical Character Recognition (OCR) designed for clean documents, STR requires robust methods to handle diverse text appearances. Standard recognizers such as CRNN [26], ASTER [27], and MORAN [19] are commonly used to evaluate STISR methods, with OCR accuracy being a key metric for assessing super-resolved images.

Current super-resolution methods often rely on pixel-domain losses (e.g., Mean Absolute Error or Root Mean Squared Error, pixel-wise), which may not correlate well with perceptual quality or semantic fidelity, especially for text. To address this, some approaches incorporate perceptual losses like VGG loss [10], focusing on visual aesthetics. Our work proposes a novel loss function inspired by TrOCR [16], capturing both visual and semantic information, aiming to improve downstream text recognition tasks.

3 Methodology

3.1 Preliminaries

Contrastive Language-Image Pre-Training (CLIP): CLIP [25] bridges the semantic gap between low-level image features and high-level concepts through pre-training on a massive dataset of image-text pairs. It employs separate image and text encoders to generate aligned embeddings, maximizing cosine similarity between matching pairs. This capability is crucial for STISR, where reconstructing semantically meaningful text from low-resolution images is essential. CLIP's proficiency in understanding textual content, even in challenging scenarios like noise and occlusions [33], strengthens its potential application in STISR.

StyleGAN: StyleGAN [11,12] is a generative adversarial network known for high-quality image generation. It introduces a style-based generator that maps latent codes to an intermediate latent space (W), which controls the generator via adaptive instance normalization (AdaIN) [8]. This separation allows fine-grained control over image features, enabling detailed texture modeling. StyleGAN2 [12] further refines this approach for superior image synthesis. StyleGAN's ability to capture intricate textures makes it suitable for generating high-resolution text images, addressing the diverse textures found in natural text settings. By combining CLIP's semantic understanding with StyleGAN's texture synthesis, we propose a novel STISR approach that overcomes current limitations.

TextZoom: Our work utilizes the TextZoom dataset [28], a comprehensive collection of real-world text images specifically designed for STISR. The dataset is constructed from SR-RAW [29] and RealSR [1], with images captured at varying focal lengths, leading to inherent misalignment and ambiguity between LR and HR pairs. TextZoom comprises 21,740 LR-HR image pairs, with 17,367 pairs for training and 4,373 for testing, divided into three subsets: Easy (1,619 samples) with minimal misalignment and ambiguity, Medium (1,411 samples) with moderate challenges, and Hard (1,343 samples) with significant misalignment and ambiguity. These variations make TextZoom an ideal and challenging test-bed for evaluating STISR algorithms. In our analysis, we utilize all subsets of the TextZoom dataset to ensure a comprehensive evaluation.

3.2 Architecture and Rationale

Our proposed architecture for STISR comprises three key modules, each playing a crucial role in the super-resolution process (Fig. 1). The first module focuses on textual understanding, the second on guided image generation, and the third on super-resolution reconstruction.

Fine-Tuning CLIP for Textual Understanding: A pre-trained CLIP model is fine-tuned on image-text pairs in the TextZoom dataset. This training process incorporates two distinct types of pairs: (1) HR images paired with their corresponding text labels, and (2) LR versions of the same images with the same text labels. This fine-tuning ensures that CLIP learns to associate both image resolutions with the same textual content. In essence, CLIP becomes adept at producing similar embeddings, numerical representations capturing essential information, for both LR and HR versions of an image as long as they contain the same text. The image encoder of this fine-tuned CLIP model serves as our "textual understanding" component. This fine-tuned CLIP is now frozen for further downstream uses.

Adapting StyleGAN2 for Guided Image Generation: Instead of using a randomly sampled latent code as input, we use the text-aware embeddings from the fine-tuned CLIP model to guide StyleGAN2's generator. This modification allows us to incorporate crucial textual information into the image generation process. Furthermore, we incorporate the LR input image at various scales to provide localized visual cues, complementing CLIP's global semantic guidance. This combined approach leads to a more faithful reconstruction during the super-resolution process. We further modify StyleGAN2 to output a set of feature maps enriched with both textual and visual information, providing a more comprehensive representation for the subsequent SR branch.

Integrating with SR Branch: Inspired by TPGSR, we use the feature maps generated by the modified StyleGAN2 as textual priors for the SR branch. These priors effectively integrate the semantic text information from CLIP and the visual details extracted from the LR image. The SR branch, based on TPGSR's SR Module, utilizes these informative priors alongside the LR image to produce the final super-resolution text image. This SR Module comprises TP-Guided SR blocks, which build upon established SR and STISR methods [15,18,28,31]. However, unlike TPGSR which relies on features from its TP transformer, our architecture directly feeds the richer feature maps produced by our modified StyleGAN2. These features are concatenated along the channel dimension and then projected back to image features, guiding the SR Module to generate a textually accurate HR output.

This combined approach allows our architecture to harness the strengths of CLIP for textual understanding, StyleGAN2 for guided image generation with textual influence, and established SR techniques for reconstructing high-quality HR images.

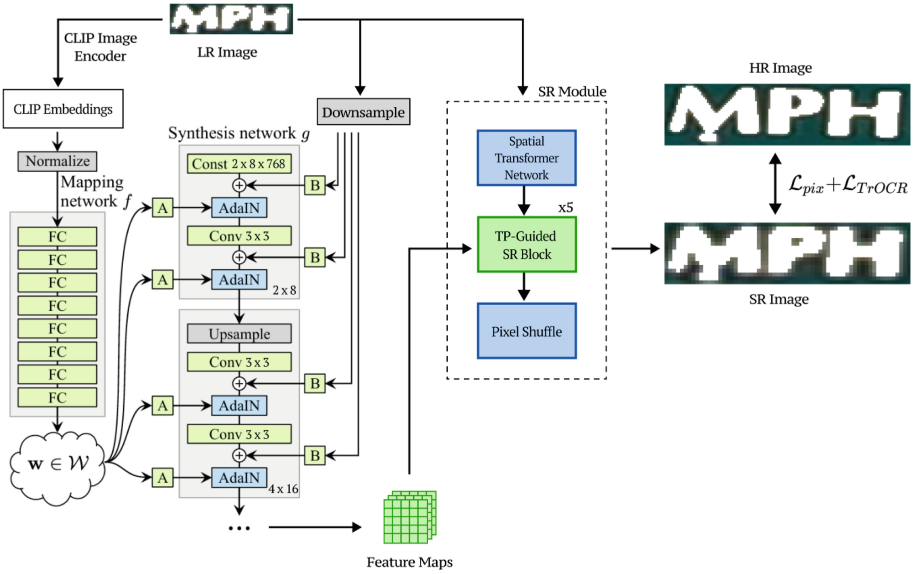


Fig. 1. Overall Architecture: The fine-tuned CLIP model provides text-aware image embeddings, guiding the modified StyleGAN2. The resulting feature maps are infused into SR branch.

3.3 Training Loss

To train our model effectively, we use a combination of pixel-level and text recognition losses. The pixel-level loss ensures the visual quality of the generated HR image ($\hat{I}_H$) compared to the ground truth HR image (I_H). As shown in Eq. 1 the pixel-level loss combines the MSE loss and the Gradient Prior Loss, which encourages sharp edges around characters and smooth backgrounds by comparing gradients in the generated and ground truth images.

$$\mathcal{L}_{pix} = \alpha||I_H - \hat{I}_H||_2^2 + \beta||\nabla I_H - \nabla \hat{I}_H||_1 \tag{1}$$

We also introduce a novel text recognition loss in Eq. 2 by utilizing an already existing pre-trained text recognition model, TrOCR [16], to guide the super-resolution process. TrOCR consists of an encoder (E_T) and a decoder (D_T). The encoder converts an image into a compact representation that captures essential information about the text, while the decoder predicts the actual characters in the image, outputting logits which are nothing but numerical scores for each possible character.

Let e_H and $\hat{e}_H$ be the encodings of the real and generated HR images, respectively, i.e., $e_H = E_T(I_H)$ and $\hat{e}_H = E_T(\hat{I}_H)$. Let d_H and $\hat{d}_H$ be the corresponding logits, i.e., $d_H = D_T(e_H)$ and $\hat{d}_H$ = $D_T(\hat{e}_H)$.

The text recognition loss in Eq. 2 encourages the generated HR image to have similar encodings and logits to the real HR image.

$$\mathcal{L}_{TrOCR} = \lambda_1||e_H - \hat{e}_H||_1 + \lambda_2||d_H - \hat{d}_H||_1 + \lambda_3 KL(d_H, \hat{d}_H) \tag{2}$$

In Eq. 2 the first two terms minimize the difference between the encodings and logits of the real and generated HR images. The third term uses the Kullback-Leibler (KL) divergence to ensure the predicted character probabilities (derived from the logits) of the real and generated HR images are similar.

The final training loss given in Eq. 3 is a combination of the pixel-level and text recognition losses that guides our model using the strengths of TrOCR.

$$\mathcal{L} = \gamma\mathcal{L}_{pix} + \delta\mathcal{L}_{TrOCR} \tag{3}$$

3.4 Evaluation Metrics

Standard image quality metrics like Peak Signal-to-Noise Ratio (PSNR) and Structural Similarity Index (SSIM) are commonly used to gauge the fidelity of reconstructed images. However, these metrics do not fully capture the effectiveness of STISR for improving text recognition.

To address this, we follow a two-step evaluation process. First, we apply different STISR methods to LR text images to generate their SR counterparts. Next, we evaluate the recognition accuracy of the generated SR images using established text recognition models like ASTER, CRNN, and MORAN. By evaluating recognition accuracy, we directly quantify the impact of STISR on downstream text understanding tasks. This ensures that STISR models are evaluated not just for visual quality but also for their ability to enable accurate text recognition in real-world scenarios.

4 Experiments and Results

4.1 Experiment Settings

Environment: We trained our models on a Google Cloud Platform instance with 8 NVIDIA V100 GPUs (16GB VRAM each), 8 CPU cores (16 virtual CPUs), and 104GB of memory. The software environment consisted of Ubuntu 20.04.6 LTS, Python 3.10.14, and PyTorch 2.3.1. We employed PyTorch's DistributedDataParallel (DDP) module, achieving faster training than the DataParallel module used in prior works.

Training Procedure: We adopted a three-stage training process to optimize model performance and stability:

1. Initial Training (500 epochs): We train the model solely on image super-resolution, excluding the TrOCR loss. This stage utilizes a larger batch size of 128 for faster training.
2. Batch Size Reduction (30 epochs): Before introducing the TrOCR loss, we reduce the batch size to 8.
3. Fine-tuning with TrOCR Loss (100 epochs): We incorporate our novel TrOCR loss and continue training with the reduced batch size. This stage refines the model to produce sharper and more textually accurate results.

Hyper-parameters: Our model uses the Adam optimizer, a widely used optimizer for deep learning models, with standard parameters $\beta_1 = 0.9$ and $\beta_2 = 0.99$. The learning rate, controlling the optimization step size, starts at 0.001 and is reduced to 0.0001 during fine-tuning with the TrOCR loss. This schedule is a common practice for initial coarse learning followed by task-specific refinement.

The loss function weights, which balance the influence of different loss components, were empirically determined. We set α to 1 and β to 0.0001 to balance the pixel-level MSE and gradient losses. For the TrOCR loss, we use equal weights ($\lambda_1 = \lambda_2 = 1$) for the encoding and logits terms and a higher weight ($\lambda_3 = 100$) for the KL divergence to emphasize similarity in predicted character probabilities. Finally, we weigh the overall pixel loss (γ) at 100 and the TrOCR loss (δ) at 5.

4.2 Fine-Tuning CLIP

We experimented with various CLIP models and fine-tuning strategies, focusing on over 30 models from OpenCLIP [9], an open-source implementation offering access to diverse pre-trained models trained on various datasets. These models encompass a range of capabilities and computational costs.

Fine-Tuning with Literal Text: We fine-tuned CLIP using paired examples of images and their corresponding literal text descriptions. To visualize the impact, we employed Relevance Maps [2], a state-of-the-art method for explaining CLIP models. These maps

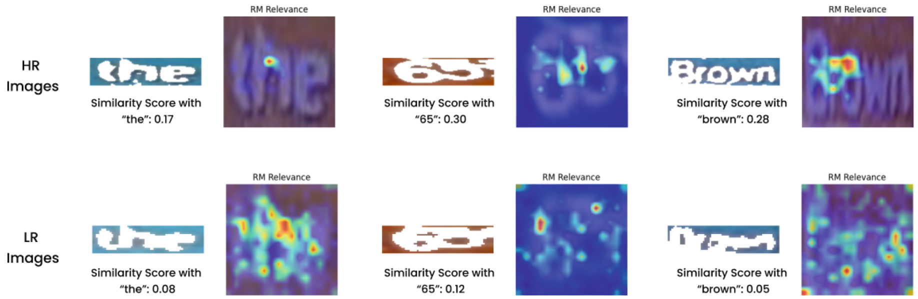

HR Images
RM Relevance
Similarity Score with "the": 0.17
RM Relevance
Similarity Score with "65": 0.30
RM Relevance
Similarity Score with "brown": 0.28
LR Images
RM Relevance
Similarity Score with "the": 0.08
RM Relevance
Similarity Score with "65": 0.12
RM Relevance
Similarity Score with "brown": 0.05

Fig. 2. Relevance Maps and similarity scores for LR and HR images before fine-tuning CLIP on literal text pairings.

highlight the parts of an image that are most influential in matching a given text description.

Figures 2 and 3 showcase these Relevance Maps alongside similarity scores for both LR and HR images before and after fine-tuning, respectively. The significant increase in similarity scores after fine-tuning demonstrates improved alignment between image and text embeddings, particularly for LR images, which approach the scores of HR images.

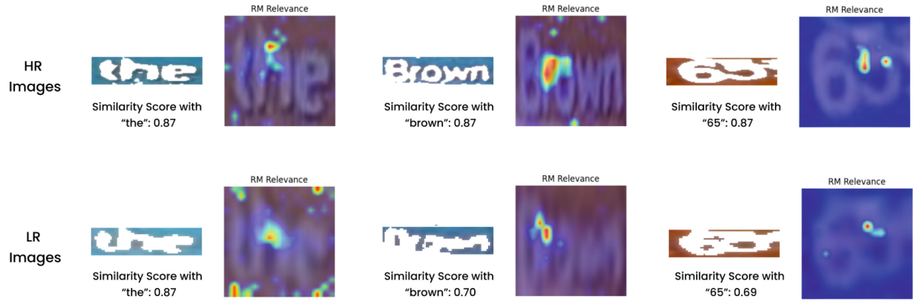

HR Images
RM Relevance
Similarity Score with "the": 0.87
RM Relevance
Similarity Score with "brown": 0.87
RM Relevance
Similarity Score with "65": 0.87
LR Images
RM Relevance
Similarity Score with "the": 0.87
RM Relevance
Similarity Score with "brown": 0.70
RM Relevance
Similarity Score with "65": 0.69

Fig. 3. Relevance Maps and similarity scores for LR and HR images after fine-tuning CLIP on literal text pairings.

A deeper analysis of the Relevance Maps reveals even more interesting details. Before fine-tuning, the attention patterns for LR and HR images differ considerably. The scattered attention on LR images suggests the model's struggle to focus on relevant textual features at lower resolutions. However, after fine-tuning, both LR and HR images exhibit focused attention on the same textual regions. This alignment signifies CLIP's ability to consistently capture semantic textual features regardless of image resolution. This successful fine-tuning equips CLIP to extract meaningful textual information even from blurry or low-quality images.

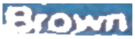
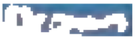
Images	Before Fine-Tuning		After Fine-Tuning	
	Similarity with "clear text"	Similarity with "blurry text"	Similarity with "clear text"	Similarity with "blurry text"
HR	0.2756	0.2813	0.9947	0.8570
LR	0.2614	0.2596	0.8624	0.9968
HR	0.2734	0.2783	0.9947	0.8576
LR	0.2614	0.2652	0.8619	0.9969
HR 	0.2738	0.2789	0.9947	0.8572
LR 	0.2686	0.2717	0.8619	0.9967

Fig. 4. Fine-tuning CLIP with Blur-Sharp Text Pairings

Fine-Tuning with Blur-Sharp Text Pairings: We further explored fine-tuning with a different approach, aiming to learn the relationship between blurry and sharp text representations directly. We used pairings of (LR image, blurry text) and (HR image, clear text) instead of literal text descriptions to fine-tune the model. Figure 4 illustrates the substantial improvements achieved in similarity scores after fine-tuning. We further investigated using this fine-tuned CLIP directly as a loss function during training. This involved calculating the similarity scores between image embeddings and text prompts "blurry" and "clear" for both SR and HR images, and backpropagating the difference as a loss. However, this approach did not outperform our proposed TrOCR loss.

4.3 Visual Comparisons

We present visual comparisons of our method's super-resolution outputs against existing techniques. Figure 5 compares the LR image, the SR outputs from the original TSRN model [28] and our TrOCR-trained TSRN. This comparison highlights the improvements achieved by incorporating our TrOCR loss into TSRN (trained with a batch size of 8 throughout).

Similarly, Fig. 6 showcases the LR image, outputs from TPGSR (Stage 1: without TrOCR loss), TrOCR-TPGSR (Stages 2 & 3: with our TrOCR loss), and the HR ground truth. This visualization emphasizes the benefits of incorporating our TrOCR loss within the TPGSR architecture. From the examples in the figure, we can observe several key improvements with our TrOCR loss. Firstly, the font style and edge clarity are significantly closer to the HR ground truth with TrOCR-TPGSR than with plain TPGSR, as illustrated by the "(Camille" and "11:00 am - 11:00" examples. Secondly, the text reconstruction quality is notably better in our case, evident from the "R" in "CALIFORNIA" example. Additionally, our approach performs better even under distortions and poor lighting conditions, as demonstrated by the "Japanese" and "SIEMENS" examples.

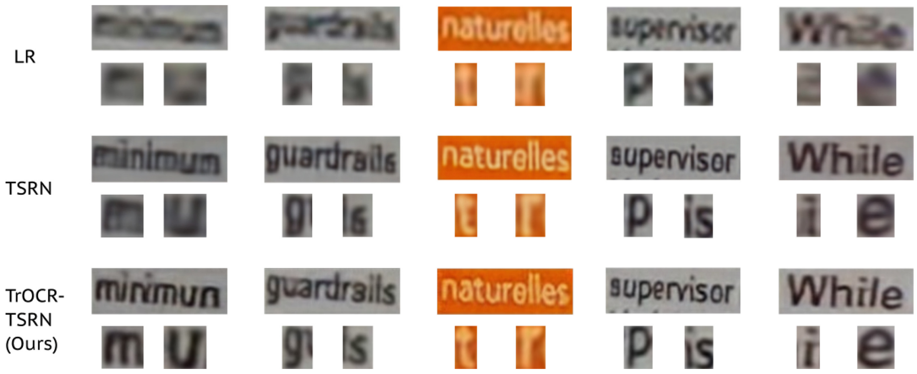


Fig. 5. Visual Comparison: LR input, SR output from the original TSRN [28] and SR output from TSRN trained with our TrOCR loss.

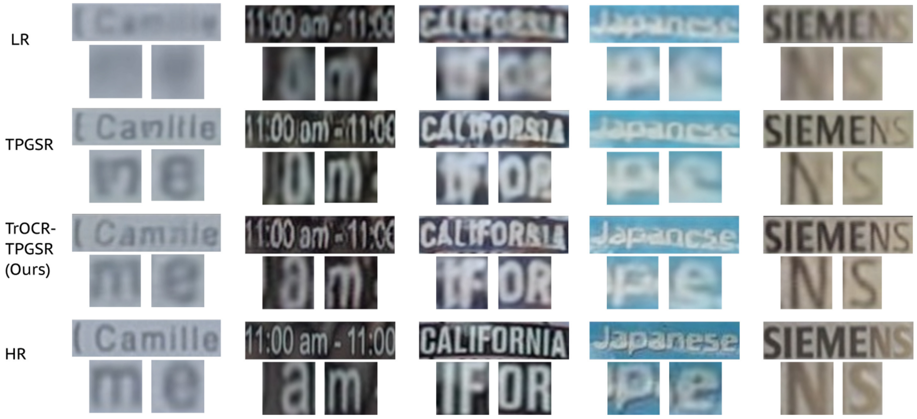


Fig. 6. Visual Comparison: LR input, TPGSR output after Stage 1 (TPGSR only), TrOCR-TPGSR output after Stage 3 (with TrOCR loss), and HR ground truth.

Figure 7 presents a comprehensive qualitative assessment of our method's performance compared to existing state-of-the-art approaches. The figure showcases the LR image alongside outputs from TrOCR-TSRN, TPGSR, TrOCR-TPGSR, our proposed method, and the HR ground truth. Notably, the visualizations reveal interesting insights into the strengths of each method. We can see that TrOCR-TPGSR excels at capturing specific font styles, as evidenced by its preservation of the blunt bend at the apex of "M" and "N" in the "SIEMENS" example, including the thicker strokes in these characters at appropriate positions. While TPGSR reconstructs the "c" and "o" well in the word "copy," our method demonstrates an advantage in reconstructing the more challenging "p" character. Additionally, TrOCR-TPGSR maintains a better font style for the letter "y" in the same word. Interestingly, when looking at the overall visual quality, our super-resolved version of words like "Diagnostic" and "Solutions" surpasses the other methods, which exhibit inconsistencies in character reconstruction. In fact, our method's output for "Solutions" appears visually superior even compared to the HR ground truth. These observations highlight the effectiveness of our proposed approach in achieving high-quality STISR.

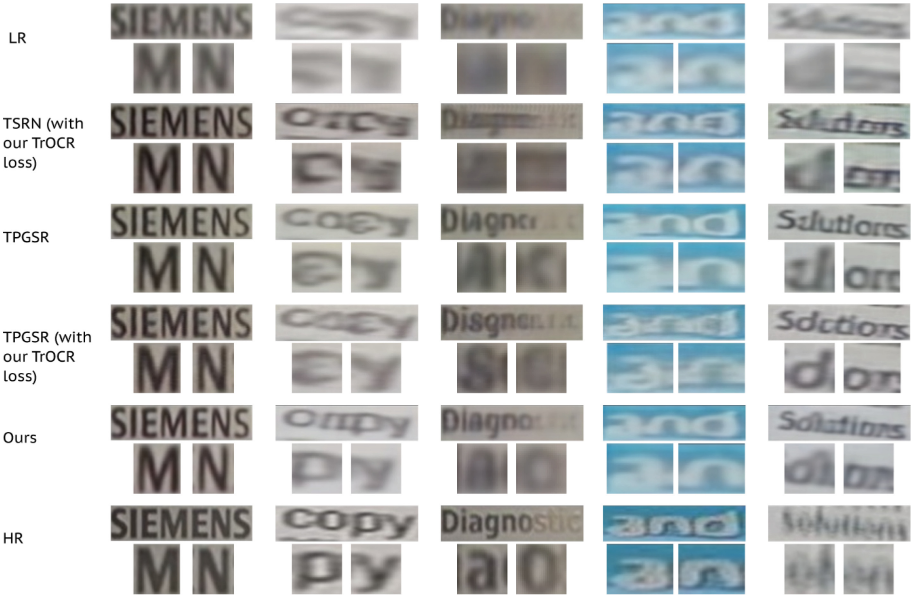


Fig. 7. Visual Comparison: LR input, outputs from TROCR-TSRN, TPGSR, TROCR-TPGSR, our method, and the HR ground truth.

4.4 Quantitative Comparision

Our method's quantitative performance is evaluated against some of the state-of-the-art STISR approaches on the TextZoom dataset. Tables 1, 2, and 3 report the recognition accuracies achieved by different methods using CRNN, ASTER, and MORAN as text recognizers, respectively. These tables present results for the easy, medium, and hard subsets of the TextZoom test set, along with the average accuracy across all subsets.

The results reveal interesting insights into the strengths of different approaches. Analyzing the CRNN recognizer's results, we observe that while plain TPGSR performs slightly better on the easy subset, our method outperforms it in the medium, hard, and overall average categories. This improvement, however, is modest at around 0.8%. Notably, the clear winner in CRNN recognizer's results is TPGSR trained with our proposed TrOCR loss, achieving an average accuracy of 50.33% and surpassing all other methods across all subsets. This suggests that, disregarding factors like font style and reconstruction details, TPGSR with TrOCR loss achieves the most accurate character reconstruction as recognized by CRNN. Additionally, TSRN also demonstrates improvement when incorporating our TrOCR loss.

Table 1. Recognition Accuracy (%) with CRNN

Method	CRNN			
	Easy	Medium	Hard	Average
BICUBIC	36.4%	21.1%	21.1%	26.8%
SRCNN	41.1%	22.3%	22.0%	29.2%
SRGAN	45.2%	32.6%	25.5%	35.1%
TSRN	52.5%	38.2%	31.4%	41.4%
TSRN (with our TrOCR loss) *	53.61%	41.32%	32.02%	43.0%
TPGSR *	56.27%	43.44%	32.02%	44.68%
TPGSR (with our TrOCR loss) *	**61.09**%	**51.10**%	**36.56**%	**50.33**%
Our Method *	55.71%	44.44%	34.25%	45.48%
HR	76.4%	75.1%	64.6%	72.4%

The results with the ASTER recognizer showcase a contrasting trend. Here, TSRN with our TrOCR loss exhibits a slight decrease in accuracy compared to plain TSRN. This could be attributed to ASTER's superior recognition capabilities, potentially allowing it to better recognize characters in the original TSRN outputs. This aligns with the observation that the performance gain for TPGSR with and without TrOCR loss is around 4% for ASTER, while it is around 6% for CRNN. Nevertheless, our method still surpasses plain TPGSR by 1.2% in ASTER recognition accuracy.

Table 2. Recognition Accuracy (%) with ASTER

Method	ASTER			
	Easy	Medium	Hard	Average
BICUBIC	67.4%	42.4%	31.2%	48.2%
SRCNN	70.6%	44.0%	31.5%	50.0%
SRGAN	69.4%	50.5%	35.7%	53.0%
TSRN	75.1%	56.3%	40.1%	58.3%
TSRN (with our TrOCR loss) *	73.13%	54.50%	39.54%	56.80%
TPGSR *	73.75%	57.90%	38.94%	57.95%
TPGSR (with our TrOCR loss) *	**77.08**%	**61.45**%	**42.96**%	**61.56**%
Our Method *	73.44%	58.68%	42.37%	59.14%
HR	94.2%	87.7%	76.2%	86.6%

Similar behavior is observed with the MORAN recognizer, suggesting that the impact of the TrOCR loss on TSRN method might vary depending on the specific text recognizer used. Overall, these quantitative results highlight the effectiveness of our proposed method in improving STISR, with TPGSR incorporating our TrOCR loss demonstrating the most significant gains in recognition accuracy irrespective of the recognizer.

Table 3. Recognition Accuracy (%) with MORAN

Method	MORAN			
	Easy	Medium	Hard	Average
BICUBIC	60.6%	37.9%	30.8%	44.1%
SRCNN	63.9%	40.0%	29.4%	45.6%
SRGAN	66.0%	47.1%	33.4%	49.9%
TSRN	70.1%	55.3%	37.9%	55.4%
TSRN (with our TrOCR loss) *	67.33%	49.75%	37.01%	52.35%
TPGSR *	68.44%	52.59%	37.08%	53.69%
TPGSR (with our TrOCR loss) *	**71.28**%	**58.54**%	**40.36**%	**57.67**%
Our Method *	68.19%	53.93%	38.64%	54.51%
HR	91.2%	85.3%	74.2%	84.1%

Table 4 compares PSNR and SSIM scores. Our method excels in PSNR but has a lower SSIM over TPGSR. Whereas TPGSR with TrOCR loss improves on both metrics. Similar to recognition results with ASTER/MORAN, TSRN with TrOCR loss sees a decrease in PSNR/SSIM. This suggests the TrOCR loss's impact on these metrics depends on the base STISR method. Overall, TPGSR with TrOCR loss achieves the highest PSNR, solidifying its pixel-level reconstruction performance, while its SSIM remains competitive with TSRN.

Table 4. Quantitative Evaluation

Method	PSNR	SSIM ($\times 10^{-2}$)
BICUBIC	20.35	69.61
SRCNN	20.78	72.28
SRGAN	21.03	73.31
TSRN	21.42	**76.91**
TSRN (with our TrOCR loss) *	21.27	76.02
TPGSR *	21.28	76.20
TPGSR (with our TrOCR loss) *	**21.68**	76.62
Our Method *	21.37	75.71

5 Conclusion and Future Work

Extracting clear and readable text from images is crucial for various applications, and STISR plays a vital role in achieving this goal. However, conventional STISR methods often fall short when dealing with the intricacies of text information within images. This limitation can manifest as blurry or poorly structured text, hindering accurate recognition. In this work, we address this by proposing a novel deep learning framework specifically designed for STISR. Our approach leverages CLIP embeddings, StyleGAN2 modifications, and a newly introduced TrOCR loss function. CLIP, with its text feature extraction strengths validated by relevance maps, strengthens our model. The effectiveness of the TrOCR loss function is further demonstrated by its ability to enhance both visual quality and recognition accuracy, even when applied to established STISR techniques. Notably, while our CLIP-StyleGAN-TrOCR model surpasses the baseline TPGSR, incorporating TrOCR loss into TPGSR yields even better results, highlighting the potential for further refinement.

While our approach demonstrates clear advancements, limitations like increased computational complexity from CLIP embeddings and potential performance drops for heavily degraded text require further exploration. Additionally, the significant performance gains from the TrOCR loss come at the cost of tenfold training time. Future work can address these limitations and explore promising avenues such as text specific perceptual losses, domain adaptation techniques for specific text domains, investigating alternative vision-language models beyond CLIP, and exploring alternative training prompts for CLIP itself.

References

1. Cai, J., Zeng, H., Yong, H., Cao, Z., Zhang, L.: Toward real-world single image super-resolution: a new benchmark and a new model. In: 2019 IEEE/CVF International Conference on Computer Vision (ICCV), pp. 3086–3095. IEEE Computer Society, Los Alamitos, CA, USA, Nov ember 2019. https://doi.org/10.1109/ICCV.2019.00318, https://doi.ieeecomputersociety.org/10.1109/ICCV.2019.00318
2. Chefer, H., Gur, S., Wolf, L.: Generic attention-model explainability for interpreting bi-modal and encoder-decoder transformers. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), pp. 397–406, October 2021
3. 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: 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 12294–12305 (2021). https://doi.org/10.1109/CVPR46437.2021.01212
4. Chen, J., Li, B., Xue, X.: Scene text telescope: text-focused scene image super-resolution. In: 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 12021–12030 (2021). https://doi.org/10.1109/CVPR46437.2021.01185
5. Chen, J., Yu, H., Ma, J., Li, B., Xue, X.: Text gestalt: stroke-aware scene text image super-resolution. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 36, pp. 285–293 (2022)
6. Dong, C., Loy, C., He, K., Tang, X.: Image super-resolution using deep convolutional networks. IEEE Trans. Pattern Anal. Mach. Intell. **38**(02), 295–307 (2016). https://doi.org/10.1109/TPAMI.2015.2439281

7. Fang, C., Zhu, Y., Liao, L., Ling, X.: Tsrgan: real-world text image super-resolution based on adversarial learning and triplet attention. Neurocomputing **455**, 88–96 (2021). https://doi.org/10.1016/j.neucom.2021.05.060. https://www.sciencedirect.com/science/article/pii/S0925231221008134
8. Huang, X., Belongie, S.: Arbitrary style transfer in real-time with adaptive instance normalization (2017)
9. Ilharco, G., Wortsman, M., Wightman, R., Gordon, C., Carlini, N., Taori, R., Dave, A., Shankar, V., Namkoong, H., Miller, J., Hajishirzi, H., Farhadi, A., Schmidt, L.: Openclip (Jul 2021). https://doi.org/10.5281/zenodo.5143773. if you use this software, please cite it as below
10. Johnson, J., Alahi, A., Fei-Fei, L.: Perceptual losses for real-time style transfer and super-resolution. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) Computer Vision - ECCV 2016. LNCS, vol. 9906, pp. 694–711. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46475-6_43
11. Karras, T., Laine, S., Aila, T.: A style-based generator architecture for generative adversarial networks (2019)
12. Karras, T., Laine, S., Aittala, M., Hellsten, J., Lehtinen, J., Aila, T.: Analyzing and improving the image quality of stylegan (2020)
13. Kim, J., Lee, J.K., Lee, K.M.: Accurate image super-resolution using very deep convolutional networks. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1646–1654 (2016). https://doi.org/10.1109/CVPR.2016.182
14. Lai, W.S., Huang, J.B., Ahuja, N., Yang, M.H.: Deep laplacian pyramid networks for fast and accurate super-resolution. In: IEEE Conferene on Computer Vision and Pattern Recognition (2017)
15. Ledig, C., et al.: Photo-realistic single image super-resolution using a generative adversarial network. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 105–114. IEEE Computer Society, Los Alamitos, CA, USA, July 2017. https://doi.org/10.1109/CVPR.2017.19. https://doi.ieeecomputersociety.org/10.1109/CVPR.2017.19
16. Li, M., et al.: Trocr: transformer-based optical character recognition with pre-trained models. In: Proceedings of the AAAI Conference on Artificial Intelligence **37**(11), 13094–13102, June 2023. https://doi.org/10.1609/aaai.v37i11.26538. https://ojs.aaai.org/index.php/AAAI/article/view/26538
17. Liang, J., Cao, J., Sun, G., Zhang, K., Van Gool, L., Timofte, R.: Swinir: image restoration using swin transformer. In: 2021 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW), pp. 1833–1844 (2021). https://doi.org/10.1109/ICCVW54120.2021.00210
18. Lim, B., Son, S., Kim, H., Nah, S., Lee, K.M.: Enhanced deep residual networks for single image super-resolution. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, July 2017
19. Luo, C., Jin, L., Sun, Z.: Moran: A multi-object rectified attention network for scene text recognition. Pattern Recogn. **90**, 109–118 (2019). https://doi.org/10.1016/j.patcog.2019.01.020. https://www.sciencedirect.com/science/article/pii/S0031320319300263
20. Luo, F., Wu, X., Guo, Y.: Functional neural networks for parametric image restoration problems. In: Ranzato, M., Beygelzimer, A., Dauphin, Y., Liang, P., Vaughan, J.W. (eds.) Advances in Neural Information Processing Systems, vol. 34, pp. 6762–6775. Curran Associates, Inc. (2021)
21. Luo, F., Wu, X., Guo, Y.: And: Adversarial neural degradation for learning blind image super-resolution. In: Oh, A., Naumann, T., Globerson, A., Saenko, K., Hardt, M., Levine, S. (eds.) Advances in Neural Information Processing Systems. vol. 36, pp. 21255–21267. Curran Associates, Inc. (2023)

22. Ma, J., Liang, Z., Zhang, L.: A text attention network for spatial deformation robust scene text image super-resolution. In: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 5901–5910. IEEE Computer Society, Los Alamitos, CA, USA, June 2022. https://doi.org/10.1109/CVPR52688.2022.00582, https://doi.ieeecomputersociety.org/10.1109/CVPR52688.2022.00582
23. Ma, J., Guo, S., Zhang, L.: Text prior guided scene text image super-resolution. IEEE Trans. Image Process. **32**, 1341–1353 (2021). https://api.semanticscholar.org/CorpusID:235669779
24. Noguchi, C., Fukuda, S., Yamanaka, M.: Scene text image super-resolution based on text-conditional diffusion models. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pp. 1485–1495 (2024)
25. Radford, A., et al.: Learning transferable visual models from natural language supervision. In: International Conference on Machine Learning (2021). https://api.semanticscholar.org/CorpusID:231591445
26. Shi, B., Bai, X., Yao, C.: An end-to-end trainable neural network for image-based sequence recognition and its application to scene text recognition. IEEE Trans. Pattern Anal. Mach. Intell. **39**(11), 2298–2304 (2017). https://doi.org/10.1109/TPAMI.2016.2646371
27. Shi, B., Yang, M., Wang, X., Lyu, P., Yao, C., Bai, X.: Aster: an attentional scene text recognizer with flexible rectification. IEEE Trans. Pattern Anal. Mach. Intell. **41**(9), 2035–2048 (2019). https://doi.org/10.1109/TPAMI.2018.2848939
28. Wang, W., Xie, E., Liu, X., Wang, W., Liang, D., Shen, C., Bai, X.: Scene text image super-resolution in the wild. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J.M. (eds.) Computer Vision - ECCV 2020. LNCS, pp. 650–666. Springer, Cham (2020)
29. Zhang, X., Chen, Q., Ng, R., Koltun, V.: Zoom to learn, learn to zoom. In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3757–3765 (2019). https://doi.org/10.1109/CVPR.2019.00388
30. Zhang, Y., Li, K., Li, K., Wang, L., Zhong, B., Fu, Y.: Image super-resolution using very deep residual channel attention networks. In: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (eds.) Computer Vision - ECCV 2018, pp. 294–310. Springer, Cham (2018)
31. Zhang, Y., Tian, Y., Kong, Y., Zhong, B., Fu, Y.: Residual dense network for image super-resolution. In: CVPR (2018)
32. Zhao, M., Wang, M., Bai, F., Li, B., Wang, J., Zhou, S.: C3-stisr: scene text image super-resolution with triple clues. In: Raedt, L.D. (ed.) Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22, pp. 1707–1713. International Joint Conferences on Artificial Intelligence Organization, August 2022. https://doi.org/10.24963/ijcai.2022/238, main Track
33. Zhao, S., Quan, R., Zhu, L., Yang, Y.: Clip4str: A simple baseline for scene text recognition with pre-trained vision-language model (2024)
34. Zhu, S., Zhao, Z., Fang, P., Xue, H.: Improving scene text image super-resolution via dual prior modulation network. ArXiv abs/2302.10414 (2023). https://api.semanticscholar.org/CorpusID:257050336

A Coverless Steganography of Face Privacy Protection with Diffusion Models

Yuan Guo and Ziqi Liu(✉)

Heilongjiang University, Harbin, China
{2023083,2232021}@s.hlju.edu.cn

Abstract. As a highly recognizable biometric feature, human face has become the first choice for identity verification. With the application of face in various important fields of society, the serious threat caused by face image information leakage has become prominent, and its privacy and security protection is becoming more and more important. Applying steganography to face images can not only effectively protect personal privacy, but also realize the secure transmission and sharing of sensitive information. Therefore, we propose a face privacy-preserving coverless steganography framework based on diffusion models. Firstly, the facial features are extracted and the feature masks are generated. Then, the DDIM sampling is used to generate the coverless stego image by combining the conditional diffusion model with the text secret key by using the generation ability of diffusion model. DDIM Inversion is used to recover the secret image with high quality. We conduct extensive experiments on CelebA-HQ and FFHQ public face datasets. Compared with the existing methods, the stego images generated and recovered by our method have higher quality and can better resist steganalysis. Our method also achieves significant advantages in terms of robustness and security, maintaining sharper recovery effects under Gaussian noise, JPEG compression, and real-world transmission. In addition, we can combine custom masks to achieve controllable local steganography, which has stronger controllability and flexibility. The proposed method can achieve a good unity of security, controllability and robustness, and is superior to the traditional steganography methods without any additional training.

Keywords: Coverless Steganography · Diffusion Model · DDIM · Privacy Protection

1 Introduction

Steganography, as a widely researched topic, aims to hide secret information within a host medium [1]. Image steganography specifically aims to covertly embed information such as images, audio, and text within a host image, meaning the host medium in image steganography is an image. The goal is to hide secret messages within the image. In typical scenarios, the sender hides the secret

A. Antonacopoulos et al. (Eds.): ICPR 2024, LNCS 15332, pp. 33–47, 2025.
https://doi.org/10.1007/978-3-031-78125-4_3

message in a cover image and transmits it to the receiver, who recovers the message. Even if the image is intercepted, no one besides the sender and receiver can detect the presence of the message [2]. Nowadays, image steganography has been widely applied in fields such as copyright protection, digital watermarking, secure information transmission, and digital forensics.

Traditional image steganography techniques often involve transforming hidden messages within the spatial or adaptive domains. Some widely used data hiding algorithms include the Least Significant Bit (LSB) method [3] and histogram-based approaches [4]. Typically, spatial domain techniques offer higher embedding capacity. With the advancement of deep neural networks, researchers have started employing autoencoder networks or invertible neural networks (INNs) [5] for data hiding, a technique known as deep steganography. The main goals of image steganography are to ensure security, preserve reconstruction quality, and improve robustness. Traditional methods typically use cover images to hide secret messages, but they often unintentionally leave behind traces of the hidden information as artifacts or local details within the carrier image. This can lead to information leakage, thereby compromising transmission security. Additionally, while these methods may achieve good reconstruction fidelity of the recovered image, they are often trained in noise-free simulated environments, rendering them vulnerable to noise, compression artifacts, and nonlinear transformations in real-world scenarios. This significantly undermines their practicality and robustness [6]. To address these challenges, recent years have seen the development of coverless data hiding methods, where secret messages are hidden without modifying the cover image. Current coverless steganography techniques frequently utilize frameworks such as CycleGAN [7] and encoder-decoder models [8], leveraging the concept of cycle consistency. Despite this, the generated container images often suffer from limited controllability, lack user-defined customization, and predominantly focus on bit-level hiding, thereby overlooking the more challenging task of embedding complete secret images.

Drawing inspiration from diffusion-based generative models, we aim to overcome the limitations of existing approaches. Research on diffusion-based generative models [9] has gained significant traction, as these models add noise to a dataset incrementally and then learn how to reverse the process, allowing for the generation of high-quality data. This method enables the production of highly accurate and detailed outputs, ranging from realistic images to coherent text sequences. The core function of these models is to gradually degrade data quality and then either restore it to its original form or transform it into a new creation. Moreover, diffusion models offer several unique attributes, such as zero-shot task performance [10], strong control over the generation process [11], natural resilience to image noise [12], and capabilities for image-to-image translation [13]. Due to their progressive denoising process, diffusion models show promising potential across various fields. The powerful control capabilities of conditional diffusion models make the generation of steganographic images highly controllable, while their generative priors ensure the visual quality of the steganographic outputs. Furthermore, diffusion models possess inherent robustness, allowing the main content of the hidden image to be retrieved even if the steganographic image is degraded during transmission.

Therefore, in this paper, we propose a face privacy protection steganography framework based on diffusion models, which aims to achieve secure, controllable and robust face privacy protection steganography. Our framework is realized by combining many properties of diffusion model, and a coverless steganography framework is implemented by using DDIM inversion [14] technique. It ensures that the hidden image has higher security and can play a more important role in information security and privacy protection.

Our contributions are summarized as follows:

(1) We propose a face image coverless steganography technique based on diffusion models, combining face feature masks and conditional diffusion models, and utilizing DDIM for inversion. Our method achieves a steganography framework specifically for face images without any additional complex training processes.
(2) We introduce the Stable Diffusion inpainting model to coverless steganography of face images, ensuring higher quality of generated steganographic and recovered images. We also achieve controllable local steganography by creating customized masks, enhancing its controllability and flexibility.
(3) Experimental results on the CelebA-HQ and FFHQ public datasets demonstrate that our method significantly outperforms existing methods in both network environments and real-world degradations, effectively resisting steganalysis while successfully achieving better reconstruction quality, higher robustness, and security.

2 Related Work

2.1 Steganography

Cover-Based Methods: Traditional Image Steganography: Traditional image steganography can be divided into two categories based on the domain where the steganography process occurs: spatial domain and frequency domain. Spatial Domain: The most popular methods include the Least Significant Bit (LSB) [3], Pixel Value Differencing (PVD) [15], and Histogram Shifting (HS) [4]. Frequency Domain: Frequently used methods include Discrete Cosine Transform (DCT) [16] and Discrete Wavelet Transform (DWT) [17]. In recent years, deep learning has been introduced into image steganography. HiDDeN [18], SteganoGAN [19], and Baluja [20] have achieved a balance between capacity, secrecy, and noise robustness, significantly improving the effective payload capacity of steganography. HiNet [21] and PRIS [22] incorporated invertible neural networks (INNs) [5] into image steganography, enabling both image hiding and recovery within a single INN model.

Coverless Methods: This emerging technology in information hiding embeds secret information without altering the cover medium. Zhou et al. [23] proposed a coverless data hiding scheme using partially repeated images. Mu and Zhou [24] used secret image copies, each sharing a similar patch with the secret image. Liu et al. [25] proposed a scheme based on DenseNet features and DWT sequence mapping. Lu et al. [26] developed a method using unsupervised learning to construct a complete basis set. Li et al. [27] proposed a method based on face fusion recognition with CNNs for encryption and decryption. Yu et al. [28] introduced a reversible image transformation technique using diffusion models, achieving better performance.

2.2 Diffusion Models

Diffusion models are currently one of the most advanced generative models, initially proposed by Sohl-Dickstein et al. in 2015 [29]. Owing to their remarkable generative capabilities, diffusion models have recently found widespread application across various image-related domains, including image generation [30], restoration [12], and translation [13]. To address the main drawback of extended training and inference times for diffusion models, numerous studies have focused on optimizing these models [11]. Recent studies have also proposed limiting the change region by using masks [31], thus retaining the background while performing meaningful image editing. "Text Inversion" [32] and "DreamBooth" [33] techniques allow users to fine-tune diffusion models by providing a few example images, enabling personalized image content generation.

3 Method

3.1 Relevant Definitions in Our Steganography

Before delving into the specifics of our method, we will first clearly define the components involved in the image steganography task, as depicted in Fig. 1. This task involves four types of images: the secret image (X_{secret}), the secret image mask (X_{mask}), the stego image (X_{stego}), and the recovered image (X_{rev}), along with two key processes: the hiding process and the revealing process. To precisely control this process, we use the FaceParsing model [34] to extract the mask X_{mask} from the secret image. This mask together with the secret image goes through the hiding process to generate the stego image X_{stego}. When the stego image is transmitted over the Internet, the quality of the image may be degraded, and a degraded stego image X'_{stego} can be obtained. Despite this, our revealing process can still recover the recovered image X_{rev} from X'_{stego} using X'_{mask}, maintaining semantic consistency of the content.

In the following sections, we will provide a detailed explanation of how to utilize the diffusion model and the FaceParsing model [34] to implement our method. Specifically: In Sect. 3.2, we will analyze the principles of the Denoising Diffusion Implicit Model (DDIM). In Sect. 3.3 we will describe in depth how to implement our face coverless steganography framework.

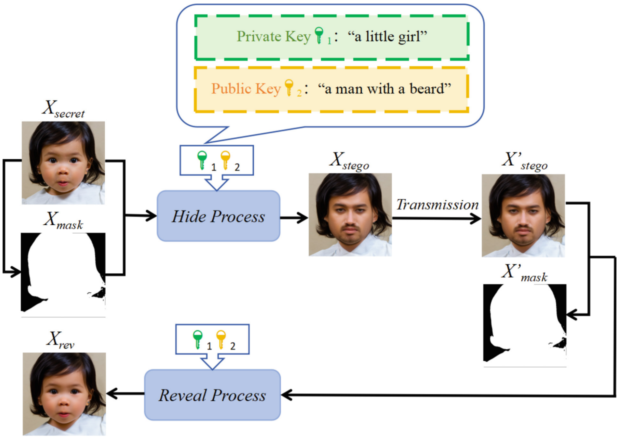


Fig. 1. The definition and composition of steganography of face image.

3.2 DDIM for Image Reversible Transformation

The DDIM is a diffusion model that utilizes deterministic inference to generate high-quality images. This model aims to improve the generation process of traditional diffusion models by reducing randomness, thereby enhancing the quality and efficiency of the generated samples.

DDIM defines its diffusion model through two main phases: the forward phase and the reverse sampling phase. In the forward phase, the model gradually adds noise to a clean image, simulating the process of the image becoming progressively distorted. Specifically, the forward process in DDIM [14] can be described by the following equations:

$$x_t = \sqrt{\alpha_t} x_{t-1} + \sqrt{1-\alpha_t}\epsilon, \quad \epsilon \sim \mathcal{N}(0,1) \tag{1}$$

where α_t is a pre-defined noise level parameter, ϵ is the random noise sampled from a standard Gaussian distribution, and x_t is the image state at time step t. The range of time step t is $[1, T]$.

In the reverse sampling phase, the model adopts the inverse process, gradually restoring the clean image by estimating and removing the noise. This process not only reduces random variations during generation but also improves the clarity and detail representation of the image, thereby generating more realistic and high-quality images. The reverse sampling process of DDIM can be described by the following equation:

$$x_s = \sqrt{\bar{\alpha}_s} f_\theta(x_t, t) + \sqrt{1-\bar{\alpha}_s-\sigma_s^2}\epsilon_\theta(x_t, t) + \sigma_s\epsilon, \quad f_\theta(x_t, t) = \frac{x_t - \sqrt{1-\bar{\alpha}_t}\epsilon_\theta(x_t, t)}{\sqrt{\bar{\alpha}_t}} \tag{2}$$

where $\epsilon \sim \mathcal{N}(0,1)$ represents Gaussian noise randomly sampled with σ_s^2 as the noise variance, and $f_\theta(., t)$ relies on a pre-trained noise estimator $\epsilon_\theta(., t)$ with $\bar{\alpha}s = \prod i = 1^t \alpha_i$. DDIM uniquely allows for non-adjacent sampling steps, meaning t and s can take any two steps where $s < t$, enhancing the flexibility and speed of the sampling process. Additionally, if set the noise variance σ_s to zero in Eq. 2, the DDIM sampling process becomes deterministic, meaning the sampling result is fully determined by the initial value x_T, which acts as a latent encoding. This deterministic process can also be described through the framework of an ordinary differential equation (ODE), where an ODE solver [14] is employed to resolve the corresponding ODE.

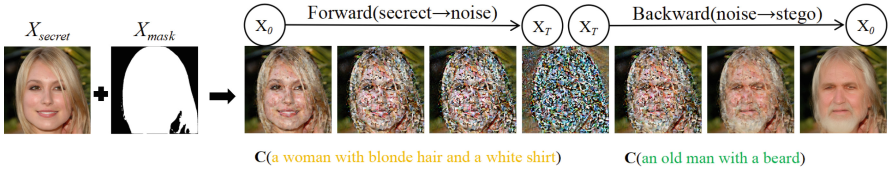


Fig. 2. We employ a conditional diffusion model for image translation across various scenarios. In this instance, we utilize two distinct prompts to transform an image of a woman into an image of an old man.

We choose to implement the diffusion model using deterministic DDIM, which not only simplifies the model's complexity but also enhances its predictability and controllability. Using the conditional diffusion model, text condition and mask are used as inputs to precisely guide the generation of image content. As illustrated in Fig. 2, our process involves transforming an image of a woman into an image of an elderly man. In this process, we begin by applying Eq. 1 to introduce noise into the woman's image during the forward sampling stage, resulting in an intermediate noise state. Next, for the backward sampling phase detailed in Eq. 2, we input a specific text condition (prompt: "an old man with a beard") to remove the noise and produce the stego image. Both the text condition (c) and mask (X_{mask}) are utilized as input conditions. The sampling process that iteratively refines the image from the noisy state (x_T) back to the clean state (x_0) is executed using the pre-trained noise estimator ϵ_θ is as follows:

$$x_0 = \text{ODESolve}(x_T, X_{mask}; \epsilon_\theta, c, T, 0) \tag{3}$$

To achieve reversible image transformation, we employ the DDIM Inversion method based on deterministic DDIM. As the name implies, this method transforms the image into potential noise and then restores it to the original image. The concept draws on the approximation of forward and backward differentials used in solving ordinary differential equations. Intuitively, for deterministic DDIM, it allows for flexibility in the steps s and t in Eq. 1, with Eq. 2 accommodating any two steps. When $s < t$, Eq. 2 performs the backward process, while when $s > t$, Eq. 2 carries out the forward process. Given the similarity in trajectories between the backward and forward processes, the input and output images are nearly identical, and the intermediate noise x_T as an effective inverted latent variable. In our research, we apply the following formulation:

$$x_T = \text{ODESolve}(x_0, X_{mask}; \epsilon_\theta, c, 0, T), \quad x'_0 = \text{ODESolve}(x_T, X'_{mask}; \epsilon_\theta, c, T, 0) \tag{4}$$

DDIM Inversion describes the transformation where the original image x_0 is converted to a latent code x_T, and subsequently, this latent code x_T is reverted back to the original image, with the output image being denoted as x'_0 and approximately equal to x_0. Using the DDIM Inversion method, we establish a reversible relationship between the image and latent noise. By utilizing the image translation framework constructed with deterministic DDIM, we can complete the entire reversible image transformation through two DDIM Inversion cycles. This technique not only serves as the core of our coverless image steganography framework but also is key to ensuring the reversibility of the steganography process. The reversibility of this method means that even in complex image processing, the integrity and accuracy of the image content can be maintained.

3.3 Face Steganography Based on Diffusion Models

Our framework is built upon a conditional diffusion model, where the noise estimator utilizes a mask and two different conditions as inputs. In our approach, these conditions function as private and public keys, denoted as K_{pri} and K_{pub} respectively. The detailed workflow is illustrated in Fig. 3. We will introduce our coverless steganography framework in two segments: the hiding process and the revealing process.

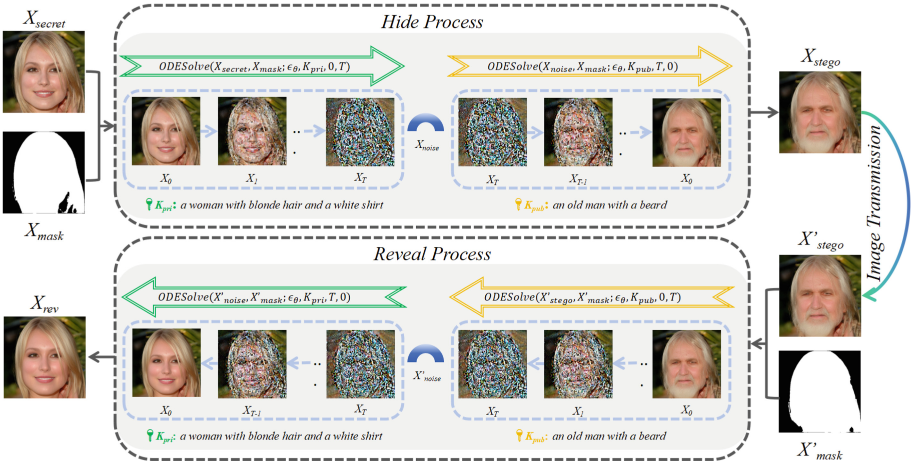

X_{secret}
X_{mask}
Hide Process
$ODESolve(X_{secret}, X_{mask}; \epsilon_\theta, K_{pri}, 0, T)$
X_0
X_1
X_T
X_{noise}
$ODESolve(X_{noise}, X_{mask}; \epsilon_\theta, K_{pub}, T, 0)$
X_T
X_{T-1}
X_0
K_{pri}: a woman with blonde hair and a white shirt
K_{pub}: an old man with a beard
X_{stego}
Image Transmission
X'_{stego}
Reveal Process
$ODESolve(X'_{noise}, X'_{mask}; \epsilon_\theta, K_{pri}, T, 0)$
$ODESolve(X'_{stego}, X'_{mask}; \epsilon_\theta, K_{pub}, 0, T)$
X_{rev}
X_0
X_{T-1}
X_T
X'_{noise}
X_T
X_1
X_0
K_{pri}: a woman with blonde hair and a white shirt
K_{pub}: an old man with a beard
X'_{mask}

Fig. 3. We opt for a conditional diffusion model that accommodates conditional inputs to steer the outcomes of generation. Furthermore, we employ deterministic DDIM as our sampling approach and utilize two distinct conditions specified by the model (K_{pri} and K_{pub}) to serve as the private and public keys, respectively.

Hiding Process: In the hiding phase, we facilitate the transformation between the secret image X_{secret} and the steganographic image X_{stego} via the deterministic DDIM's forward and backward processes. To ensure variability in the images pre- and post-transformation, we engage the pre-trained conditional diffusion model with differing conditions for each process. These conditions also serve dual roles as private and public keys (K_{pri} and K_{pub}). Specifically, we use a generated mask X_{mask} from the original secret image to control the depiction of people independently from the background and other elements, employing K_{pri} in the forward process and K_{pub} in the reverse. The resulting steganographic image X_{stego} is then sent across the Internet, accessible to all potential recipients. This setup hinges on the effectiveness of the conditions: the private key outlines the content of the secret image, while the public key influences the steganographic image's content. In this model, the public key is inferable from the steganographic image itself, thus, it need not be transmitted separately. Conversely, the private key is crucial for accurate image recovery and must remain confidential.

Revealing Process: In the revealing phase, we assume the steganographic image X'_{stego} has been transmitted online and possibly altered. The recipient utilizes the same conditional diffusion process with the corresponding keys, employing a reverse sequence to the hiding process, to restore the original secret image. This involves regenerating a control mask from the steganographic image X'_{stego}, now called X'_{mask}, using K_{pri} in the forward process. Unlike the hiding phase, where K_{pub} is used forward and K_{pri} backward, the revealing phase adjusts these roles. This method of coverless image steganography doesn't require training or fine-tuning the diffusion model specifically for steganography tasks; rather, it leverages the inherent reversible image transformation capabilities of DDIM Inversion. The forthcoming section will delve into this framework's specific applications and operational details, demonstrating its efficacy in safeguarding the privacy and security of image content in real-world scenarios.

4 Experimental Results

4.1 Implementation Details and Setup

Experimental Settings: In our experiments, we utilized the FaceParsing model to generate facial masks and chose Stable Diffusion V2-Inpainting, provided by Huggingface, as our conditional diffusion model. We used deterministic DDIM inversion to perform the inversion, with both forward and reverse processes comprising 50 steps each. To facilitate reversible image transformation, we adjusted the guidance scale to 1.0 and set the strength to 0.99.

Data Preparation: We used two facial image datasets, CelebA-HQ and FFHQ. CelebA-HQ contains 30,000 high-resolution facial images, and FFHQ has 70,000 high-definition images at 1024 × 1024 resolution. From these, we curated 240 images, named StegFace240. We used the BLIP [13] model to generate descriptive textual information for the images as the private key, with the public key manually modified. To validate our method, we compared it against several state-of-the-art image steganography techniques, demonstrating its effectiveness. Our method requires no training, and all experiments were conducted using a GeForce RTX 3090 GPU card.

4.2 Comparison with SOTA Methods

In our experiments, we compared our method with various techniques on the StegFace240 dataset. Considering that the application of diffusion models in image steganography is relatively novel, we implemented several versions of the Stable Diffusion model, including SDXL, SDXL-Inpt, and SD-Inpt. As shown in Fig. 4, we compared the quality of steganographic and recovered images generated by different methods. It is evident that the steganographic images produced by our method efficiently conceal the secret images without introducing noticeable artifacts or unrealistic details, making anomalies virtually undetectable to the human eye. Moreover, our steganographic images support seamless modifications of facial features such as gender, age, and beard, with high controllability. Regarding controllability (Fig. 5), our approach enables steganography in targeted regions while leaving other areas unaffected. It ensures the accurate preservation of the secret image's semantic information using the private key, thereby exhibiting outstanding fidelity.

Our method not only allows for highly accurate recovery of the secret image but also minimizes the difference between the original and recovered images. We adopt four different metrics to evaluate the quality of the secret image and the recovered image, including PSNR, SSIM, LPIPS and FID as shown in Table 1. Higher PSNR and SSIM scores indicate better quality of the recovered images, while lower LPIPS and FID scores suggest that the generated images are closer to real images in terms of visual perception and style, reducing the likelihood of being detected as containing steganographic information. Additionally, we used Face++ and Aliyun's facial recognition models to verify the recovery effectiveness (Table 1). The facial recognition rate between the recovered images and the secret images achieved over 90% on both models, attaining the highest confidence levels. The results show that our method significantly outperforms other methods across all metrics.

4.3 Steganalysis

To evaluate the security of the steganographic images, we employed both traditional statistical methods and deep learning-based steganalysis techniques to determine whether the images can withstand detection by existing steganalysis tools. As shown in the left side of Fig. 6, we used the open-source steganalysis tool StegExpose [35] to test the anti-steganalysis capability of our model. By adjusting different detection thresholds, we generated ROC curves. The closer the area under the ROC curve is to 0.5, the closer the detection accuracy is

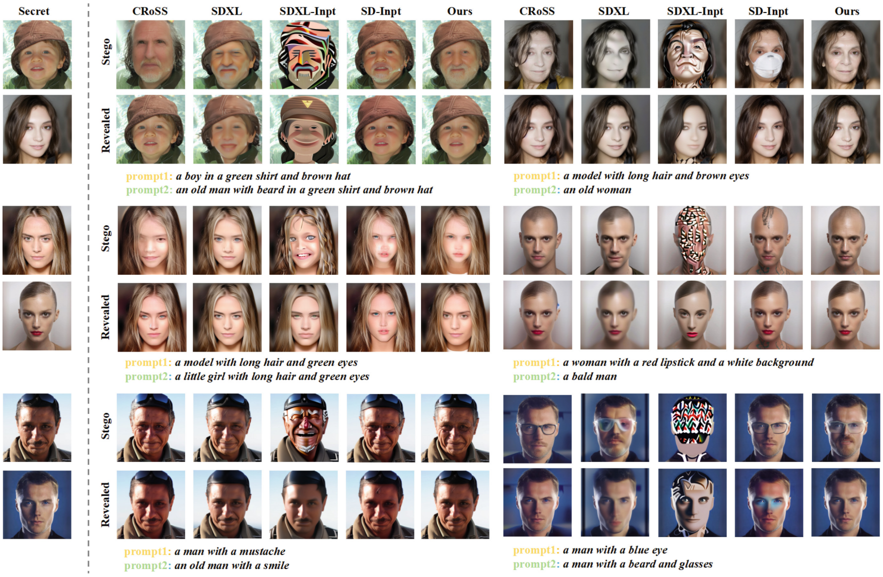

Secret
CRoSS
SDXL
SDXL-Inpt
SD-Inpt
Ours
CRoSS
SDXL
SDXL-Inpt
SD-Inpt
Ours
Stego
Revealed
prompt1: a boy in a green shirt and brown hat
prompt2: an old man with beard in a green shirt and brown hat
prompt1: a model with long hair and brown eyes
prompt2: an old woman
Stego
Revealed
prompt1: a model with long hair and green eyes
prompt2: a little girl with long hair and green eyes
prompt1: a woman with a red lipstick and a white background
prompt2: a bald man
Stego
Revealed
prompt1: a man with a mustache
prompt2: an old man with a smile
prompt1: a man with a blue eye
prompt2: a man with a beard and glasses

Fig. 4. Our method compares with other methods for steganography and image recovery. It can be seen that the steganographic image generated by our method has high visual quality, is not easy to be detected, and the recovered image has a high similarity to the secret image.

to random guessing, indicating better resistance to steganalysis detection. The results clearly show that our method exhibits low detection accuracy, suggesting that the steganographic images generated by our model possess high security and can effectively deceive the StegExpose tool.

In the right side of Fig. 6, we used the deep learning-based steganalysis tool SRNet [36] and tested the steganographic images produced by various methods using the StegFace240 dataset. We retrained SRNet by gradually increasing the number of steganographic images used for training. The data in the figure indicates that compared to other methods, our proposed method shows significantly lower detection accuracy, further demonstrating the strong anti-steganalysis capability of our method. Table 2 presents the detection accuracy of different image hiding methods using SRNet. Ideally, the closer the detection accuracy is to 50%, the better the performance of the image hiding algorithm. Our method achieved a detection accuracy of 55.25%, indicating that the steganographic images are almost impossible to accurately detect as containing hidden information.

4.4 Robustness Analysis

To assess the robustness of our method, we performed a series of simulated degradation experiments, including the addition of Gaussian noise and JPEG compres-

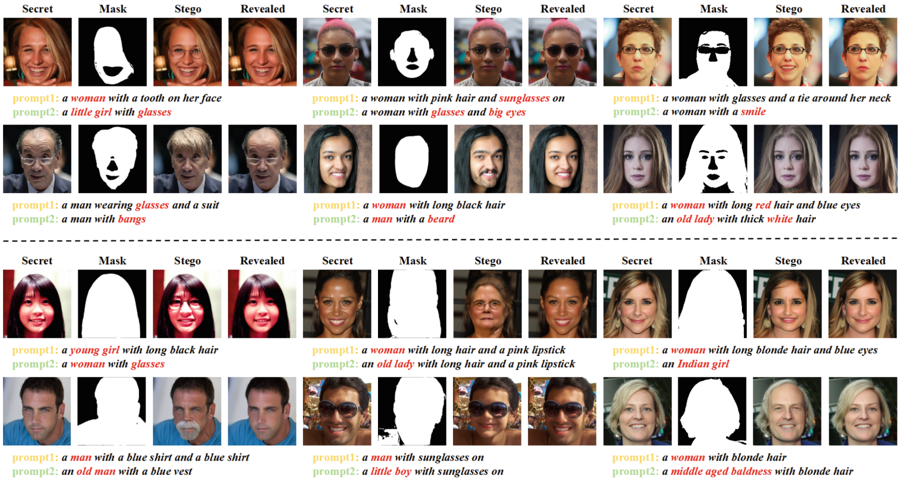

Secret
Mask
Stego
Revealed
prompt1: a woman with a tooth on her face
prompt2: a little girl with glasses
prompt1: a woman with pink hair and sunglasses on
prompt2: a woman with glasses and big eyes
prompt1: a woman with glasses and a tie around her neck
prompt2: a woman with a smile
prompt1: a man wearing glasses and a suit
prompt2: a man with bangs
prompt1: a woman with long black hair
prompt2: a man with a beard
prompt1: a woman with long red hair and blue eyes
prompt2: an old lady with thick white hair
prompt1: a young girl with long black hair
prompt2: a woman with glasses
prompt1: a woman with long hair and a pink lipstick
prompt2: an old lady with long hair and a pink lipstick
prompt1: a woman with long blonde hair and blue eyes
prompt2: an Indian girl
prompt1: a man with a blue shirt and a blue shirt
prompt2: an old man with a blue vest
prompt1: a man with sunglasses on
prompt2: a little boy with sunglasses on
prompt1: a woman with blonde hair
prompt2: a middle aged baldness with blonde hair

Fig. 5. Our method, combined with custom face mask control, realizes controllable local steganography and has good semantic consistency.

Table 1. Comparison results of our proposed method and other methods on the StegFace240 dataset. The best results are highlighted in bold.

Methods	Secret/Reverse					
	PSNR↑	SSIM↑	LPIPS↓	FID↓	Face++↑	Aliyun↑
CRoSS	23.79	0.74	0.18	48.85	89.20	70.85
SDXL	24.56	0.75	0.31	71.71	86.02	63.14
SDXL-Inpt	19.82	0.65	0.33	117.94	68.37	32.48
SD-Inpt	26.38	0.78	0.11	30.62	93.86	81.48
Ours	**28.76**	**0.82**	**0.08**	**21.82**	**96.21**	**90.36**

sion. As illustrated in Table 3, our method exhibited remarkable adaptability to different levels of degradation, with minimal performance decline. Notably, in the presence of Gaussian noise and JPEG compression, our method achieved the highest PSNR values. Even under severe conditions such as Gaussian noise with $\sigma = 30$ and JPEG compression with QF = 20, the PSNR values remained above 20dB and 25dB respectively, whereas other methods exhibited a significant drop in fidelity.

To further prove the robustness of our method, we tested real-world degradation scenarios. We conducted steganographic image transmission and reception experiments via the WeChat network to simulate the effects of network transmission. As illustrated in Fig. 7, under such complex degradation conditions, all other methods either failed entirely or exhibited significant color distortions. In contrast, our method not only successfully revealed the general content of the

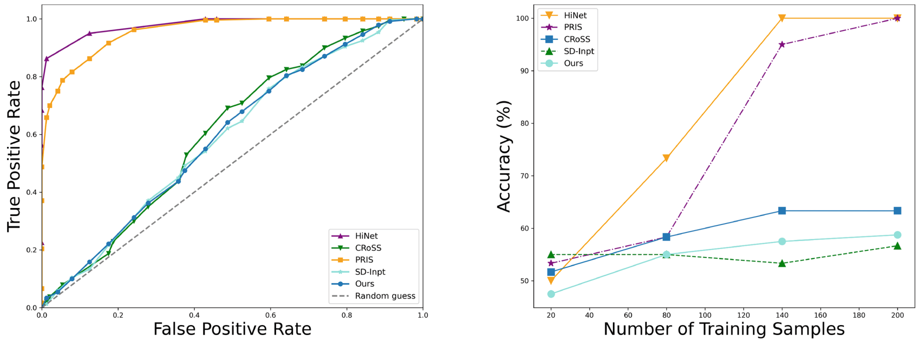


Fig. 6. The left is the ROC curves generated by different methods under the StegExpose detector. The closer the area under the curve is to 0.5, the better the method is at ideally evading the detector. The right is the results of steganalysis using SRNet. The slower the curve grows and the closer the accuracy is to 50%, the higher the method's resistance to steganalysis.

Table 2. Detection accuracy of different methods on SRNet. The best results are highlighted in bold.

Methods	Accuracy (%)±std
HiNet	77.17±0.251
PRIS	74.33±0.219
CRoSSt	57.50±0.059
SD-Inpt	**53.50±0.023**
Ours	55.25±0.049

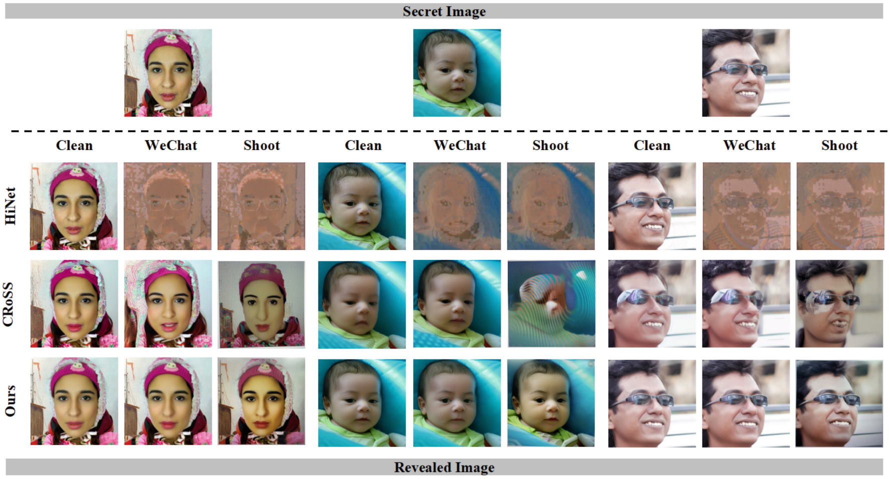


Fig. 7. In real-world scenarios, when subjected to visual downgrades under conditions labeled "Shoot" and "WeChat," our method effectively reconstructs the contents of a secret image, whereas other methods display significant color distortion or fail entirely.

Table 3. Comparison of PSNR (dB) results for our proposed method and other techniques under various levels of degradation. The best results are highlighted in bold.

Methods	Gaussian Noise			JPEG		
	$\sigma = 10$	$\sigma = 20$	$\sigma = 30$	QF = 20	QF = 40	QF = 80
HiNet	20.45	13.55	10.22	11.06	11.12	12.75
PRIS	23.83	18.29	14.90	12.86	13.02	15.66
CRoSS	20.78	19.10	17.49	20.73	21.36	22.96
SD-Inpt	24.04	21.40	19.65	24.16	25.09	25.79
Ours	**25.96**	**23.99**	**22.48**	**25.49**	**26.87**	**28.15**

secret image but also maintained good semantic consistency with the private key, once again proving the significant superiority of our approach. Compared to the latest methods, our proposed method also successfully maintained higher reconstruction quality. These experimental results fully validate the efficiency and robustness of our method under various experimental and real-world conditions.

5 Conclusion

We proposed a face privacy protection steganography framework based on diffusion models. This framework combines mask extraction models, conditional diffusion models, and deterministic DDIM techniques, which leverage the unique advantages of diffusion models to achieve coverless steganography and is difficult to detect with steganalysis tools. A large number of experiments show that compared with the existing techniques, our method has obvious advantages in the process of steganography and restoration. Moreover, the generated steganography images are diverse. Our method achieves a good balance in terms of security, controllability, and robustness.

In the future, image steganography based on diffusion models has tremendous potential for development. Continued exploration of new methods to enhance the steganographic capabilities of diffusion models, particularly in improving the ability to hide multiple pieces of information and achieving pixel-level fidelity, will have important application value. We look forward to future research addressing these limitations and further optimizing and refining the technical architecture in this field.

Acknowledgements. This work was supported partly by National Natural Science Foundation of China(62473278), Heilongjiang Province Natural Science Foundation (LH2021F056), and Heilongjiang Provincial Education Department, Grant/Award (1355091130).

References

1. Kessler, G.C., Hosmer, C.: An overview of steganography. Adv. Comput. **83**, 51–107 (2011)
2. Chanu, Y. J., Singh, K. M., Tuithung, T.: Image steganography and steganalysis: a survey. In: International Joint Conference on Artificial Intelligence (IJCAI) 52(2) (2012)
3. Chan, C.K., Cheng, L.M.: Hiding data in images by simple LSB substitution. Pattern Recogn. **37**(3), 469–474 (2004)
4. Li, Z., Chen, X., Pan, X., Zeng, X.: Lossless data hiding scheme based on adjacent pixel difference. In 2009 International Conference on computer engineering and technology, vol. 1, pp. 588–592. IEEE (2009)
5. Dinh, L., Krueger, D., Bengio, Y.: Nice: non-linear independent components estimation. arXiv preprint arXiv:1410.8516 (2014)
6. Qin, J., Luo, Y., Xiang, X., Tan, Y., Huang, H.: Coverless image steganography: a survey **7**, 171372–171394 (2019)
7. 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)
8. Zhou, Z., Sun, H., Harit, R., Chen, X., Sun, X.: Coverless image steganography without embedding. In: Cloud Computing and Security: First International Conference, pp. 123–132. Springer, China (2015)
9. 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)
10. Kawar, B., Elad, M., Ermon, S., Song, J.: Denoising diffusion restoration models. Adv. Neural. Inf. Process. Syst. **35**, 23593–23606 (2022)
11. Mou, C., Wang, X., Xie, L., Wu, Y., Zhang, J., Qi, Z., Shan, Y.: T2i-adapter: learning adapters to dig out more controllable ability for text-to-image diffusion models. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 38, No. 5, pp. 4296–4304, March 2024
12. Hertz, A., Mokady, R., Tenenbaum, J., Aberman, K., Pritch, Y., Cohen-Or, D.: Prompt-to-prompt image editing with cross attention control. https://arxiv.org/abs/2208.01626 (2022)
13. 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)
14. Song, J., Meng, C., Ermon, S.: Denoising diffusion implicit models. arXiv preprint arXiv:2010.02502 (2020)
15. Pan, F., Li, J., Yang, X.: Image steganography method based on PVD and modulus function. In: 2011 International Conference on Electronics, Communications and Control (ICECC), pp. 282–284. IEEE (2011)
16. McKeon, R.T.: Strange Fourier steganography in movies. In: 2007 IEEE International Conference on Electro/Information Technology, pp. 178-182. IEEE (2007)
17. Hsieh, M.S., Tseng, D.C., Huang, Y.H.: Hiding digital watermarks using multiresolution wavelet transform. IEEE Trans. Industr. Electron. **48**(5), 875–882 (2001)
18. Zhu, J., Kaplan, R., Johnson, J., Fei-Fei, L.: Hidden: Hiding data with deep networks. In: Proceedings of the European conference on computer vision (ECCV), pp. 657–672 (2018)

19. Zhang, K.A., Cuesta-Infante, A., Xu, L., Veeramachaneni, K.: SteganoGAN: high capacity image steganography with GANs. arXiv preprint arXiv:1901.03892 (2019)
20. Baluja, S.: Hiding images in plain sight: deep steganography. Advances in neural information processing systems, 30 (2017)
21. Jing, J., Deng, X., Xu, M., Wang, J.,Guan, Z.: Hinet: deep image hiding by invertible network. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 4733–4742 (2021)
22. Yang, H., Xu, Y., Liu, X., Ma, X.: PRIS: practical robust invertible network for image steganography. Eng. Appl. Artif. Intell. **133**, 108419 (2024)
23. Zhou, Z., Mu, Y., Wu, Q.J.: Coverless image steganography using partial-duplicate image retrieval. Soft. Comput. **23**(13), 4927–4938 (2019)
24. Mu, Y., Zhou, Z.: Visual vocabulary tree-based partial-duplicate image retrieval for coverless image steganography. Int. J. High Perform. Comput. Networking **14**(3), 333–341 (2019)
25. Liu, Q., Xiang, X., Qin, J., Tan, Y., Tan, J., Luo, Y.: Coverless steganography based on image retrieval of DenseNet features and DWT sequence mapping. Knowl.-Based Syst. **192**, 105375 (2020)
26. Lu, J., Ni, J., Li, L., Luo, T., Chang, C.: A coverless information hiding method based on constructing a complete grouped basis with unsupervised learning **6**(1), 29–39 (2021)
27. Li, Y.H., Chang, C.C., Su, G.D., Yang, K.L., Aslam, M.S., Liu, Y.: Coverless image steganography using morphed face recognition based on convolutional neural network. EURASIP J. Wirel. Commun. Netw. **2022**(1), 28 (2022)
28. Yu, J., Zhang, X., Xu, Y., Zhang, J.: Cross: diffusion model makes controllable, robust and secure image steganography. Advances in Neural Information Processing Systems **36** (2024)
29. 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)
30. 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)
31. 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)
32. Gal, R., et al.: An image is worth one word: Personalizing text-to-image generation using textual inversion. arXiv preprint arXiv:2208.01618 (2022)
33. 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)
34. Luo, L., Xue, D., Feng, X.: Ehanet: an effective hierarchical aggregation network for face parsing. Appl. Sci. **10**(9), 3135 (2020)
35. Stegexpose: a tool for detecting LSB steganography. arXiv preprint arXiv:1410.6656 (2014)
36. Corley, I., Lwowski, J., Hoffman, J.: Destruction of image steganography using generative adversarial networks. arXiv preprint arXiv:1912.10070 (2019)

CLIP-AGIQA: Boosting the Performance of AI-Generated Image Quality Assessment with CLIP

Zhenchen Tang, Zichuan Wang, Bo Peng(✉), and Jing Dong(✉)

New Laboratory of Pattern Recognition, Institute of Automation,
Chinese Academy of Sciences, Beijing, China
{tangzhenchen2024,wangzichuan2024}@ia.ac.cn,
{bo.peng,jdong}@nlpr.ia.ac.cn

Abstract. With the rapid development of generative technologies, AI-Generated Images (AIGIs) have been widely applied in various aspects of daily life. However, due to the immaturity of the technology, the quality of the generated images varies, so it is important to develop quality assessment techniques for the generated images. Although some models have been proposed to assess the quality of generated images, they are inadequate when faced with the ever-increasing and diverse categories of generated images. Consequently, the development of more advanced and effective models for evaluating the quality of generated images is urgently needed. Recent research has explored the significant potential of the visual language model CLIP in image quality assessment, finding that it performs well in evaluating the quality of natural images. However, its application to generated images has not been thoroughly investigated. In this paper, we build on this idea and further explore the potential of CLIP in evaluating the quality of generated images. We design *CLIP-AGIQA*, a CLIP-based regression model for quality assessment of generated images, leveraging rich visual and textual knowledge encapsulated in CLIP. Particularly, we implement multi-category learnable prompts to fully utilize the textual knowledge in CLIP for quality assessment. Extensive experiments on several generated image quality assessment benchmarks, including AGIQA-3K and AIGCIQA2023, demonstrate that *CLIP-AGIQA* outperforms existing IQA models, achieving excellent results in evaluating the quality of generated images.

Keywords: AI-Generated Images · CLIP · Perceptual Quality

1 Introduction

With the rapid development of generative technologies, Artificial Intelligence Generated Images (AIGIs) have become increasingly ubiquitous in modern society. From avatar generation on social media to visual effects production in movies and television, and even content creation in virtual and augmented reality, generative technologies has become an integral part of our daily experiences. However,

Z. Tang and Z. Wang: These authors contributed equally.

A. Antonacopoulos et al. (Eds.): ICPR 2024, LNCS 15332, pp. 48–61, 2025.
https://doi.org/10.1007/978-3-031-78125-4_4

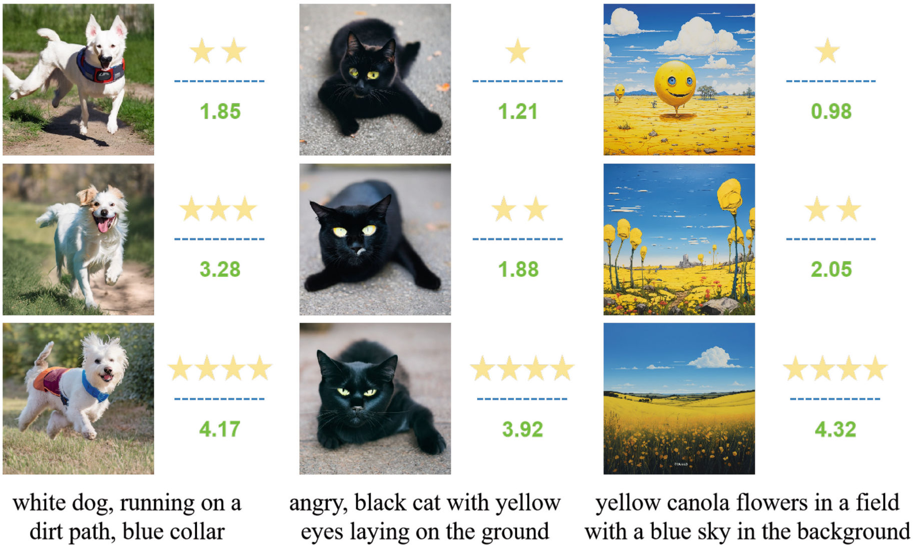

★★
1.85
★
1.21
★
0.98
★★★
3.28
★★
1.88
★★
2.05
★★★★
4.17
★★★★
3.92
★★★★
4.32
white dog, running on a dirt path, blue collar
angry, black cat with yellow eyes laying on the ground
yellow canola flowers in a field with a blue sky in the background

Fig. 1. Performance of *CLIP-AGIQA*. The star icons represent human ratings, and the green scores below the dashed line represent the scores predicted by our model. (Color figure online)

alongside these technological advancements, assessing the quality of generated images has become an emerging issue. Due to the immaturity of the technology, the quality of generated images is uneven, which can lead to unsatisfactory user experiences in some applications [14]. Therefore, developing techniques to effectively evaluate the quality of generated images is particularly important.

Quality assessment of generated images involves evaluating various dimensions through subjective and objective methods, such as the perceptual quality and the content accuracy with respect to input prompts. Recent efforts have focused on creating comprehensive databases for subjective quality assessment based on human perception and developing approaches to enhance evaluation performance [12,14,29]. Despite these advancements, existing methods struggle to keep pace with the increasing diversity of generated images. For instance, in the field of text-to-image (T2I) generative models alone, there have been at least 20 representative T2I AGI models up to 2023, as indicated by recent statistics [2,34]. Therefore, more research is needed to meet the quality assessment demands in this field.

Recent research has begun to explore CLIP's [18] (Contrastive Language-Image Pre-training) potential in image quality assessment, revealing its effectiveness in evaluating natural images [24]. CLIP demonstrates strong performance across various visual and multimodal tasks due to its extensive pre-training on language-image data. However, since CLIP is pre-trained on natural images, it may have problems to model the quality distribution of generated images effectively, leaving a gap in this area. To address this, we propose *CLIP-AGIQA*, a

CLIP-based regression model that leverages CLIP's comprehensive visual and textual knowledge to evaluate the quality of generated images. First, we design various prompts representing different quality levels to input into CLIP's text encoder, mitigating semantic ambiguities. Second, by introducing a learnable prompts strategy and utilizing multiple quality-related auxiliary prompts, we make full use of CLIP's textual knowledge. Last, our regression network then maps CLIP features to quality scores, effectively adapting CLIP's capabilities to the task of generated image quality assessment, thereby enhancing the model's performance. The specific performance of our *CLIP-AGIQA* can be seen in Fig. 1.

In summary, our primary contributions include:

- We propose *CLIP-AGIQA*, adapting the CLIP model to the task of evaluating generated image quality;
- We introduce a learnable prompts strategy and design multiple prompts of varying quality levels to fully utilize CLIP's textual knowledge for assisting in evaluating generated image quality;
- We conduct experiments on several benchmarks for generated image quality assessment such as AGIQA-3K and AIGCIQA2023, achieving state-of-the-art performance.

2 Related Work

2.1 Image Quality Assessment

Traditional image quality assessment aims to evaluate the quality of natural images, including aspects like noise, blur, compression artifacts, etc. [3]. It is categorized into three types: full-reference, reduced-reference, and no-reference. Full-reference methods compare the original and test images, commonly using metrics like PSNR and SSIM [26]. Reduced-reference methods utilize partial information from a reference image, such as RRED [21] and OSVP [27]. No-reference methods directly assess image quality using machine learning and deep learning techniques, such as BRISQUE [16], IQA-CNN [9] and RankIQA [15].

In recent years, with the development of generative technologies, assessing the quality of generated images has become increasingly important. Due to potential abnormal distortions or unrealistic structures in generated images, evaluation focuses on visual perception, including authenticity, naturalness, and coherence. Common metrics include Inception Score (IS) for assessing image quality and diversity based on classification results and KL divergence [19], Fréchet Inception Distance (FID) for evaluating visual quality by comparing feature distributions of real and generated images [7], and CLIP Score, which assesses image quality based on similarity between generated images and textual descriptions [6].

Recently, datasets like AGIQA-3K [14] and PKU-I2IQA [33] have been proposed to facilitate benchmark experiments for IQA models, focusing on the quality assessment of generated images. AGIQA-3K provides a comprehensive and diverse subjective quality database covering various generated images from

GAN, autoregressive, and diffusion models. PKU-I2IQA, the first image-to-image AIGC quality assessment database based on human perception, also conducts benchmark experiments on different IQA models. Additionally, models such as ImageReward [29] and HPS [28] construct datasets for generated images from the perspective of human preferences and proposed corresponding evaluation models, providing a benchmark for quality assessment in terms of human preferences for generated images. Despite these advancements, there remains a scarcity of specialized models for assessing the quality of generated images, necessitating further research to advance this field.

2.2 CLIP-Based Methods

CLIP [18] is a large-scale vision-language pretrained model that leverages contrastive learning to achieve cross-modal knowledge understanding. It has demonstrated strong transfer capabilities across various visual tasks such as semantic segmentation (LSeg [13]), object detection (ViLD [4]), and image generation (CLIPasso [23]).

CLIP-IQA [24] is the first work to explore CLIP in image quality assessment tasks, demonstrating that CLIP can be effectively extended to image quality evaluation. Due to the significant impact of linguistic ambiguity in quality assessment tasks [11], phrases such as “a rich image” can be particularly problematic. This phrase could either refer to an image with rich content or an image associated with wealth. CLIP-IQA design an antonym prompt strategy to leverage CLIP’s prior knowledge. However, due to the limited variety of prompts, this approach can result in inaccurate quality predictions. Moreover, this work only explored the performance of CLIP in natural image quality assessment tasks and did not address generated images. Building on this idea, we further investigate the performance of CLIP in evaluating the quality of generated images and propose a CLIP-based quality assessment regression model. By simultaneously fine-tuning our designed multi-class learnable prompts and the regression network added after CLIP, we achieve superior performance in assessing the quality of generated images.

Notably, recent methods [8,10,36] also explore CLIP for IQA, with many focusing on aesthetic evaluation. These methods stand out for their pioneering efforts in multi-modality integration for low-level vision and their impressive zero-shot performance. However, since CLIP is pre-trained on natural image-text pairs, directly using CLIP in a zero-shot manner to evaluate the quality of generated images, as done in the aforementioned methods, does not yield optimal results. Therefore, we train a CLIP-based model using generated images to better model the quality distribution of generated images.

3 Methodology

In this section, we first formalize the paradigm of a typical IQA model. Then, we provide a detailed description of the various designs we implement to adapt CLIP to the task of generative image quality assessment in *CLIP-AGIQA*.

3.1 Preliminary on IQA Models

Given an image I, a typical IQA model uses a visual encoder $V(\cdot)$ to extract visual features, followed by a regression model $R(\cdot)$ to predict the quality score. This process can be represented as follows:

$$S = R(V(I)) \tag{1}$$

In CLIP-IQA [24], only the visual encoder $V(\cdot)$ is used to extract visual features, and then an antonym prompt strategy is employed to compute the cosine similarity with the visual features to predict the quality score. Specifically, CLIP-IQA adopts antonym prompts (e.g., "Good photo." and "Bad photo.") as a pair for each prediction. Let x represent the features from the image, and t_1 and t_2 be the features from the two prompts with opposite meanings. The cosine similarity is computed as follows [24]:

$$s_i = \frac{x \cdot t_i}{||x|| \cdot ||t_i||}, \quad i \in \{1, 2\}, \tag{2}$$

and Softmax is used to compute the final score $\bar{s} \in [0, 1]$:

$$\bar{s} = \frac{e^{s_1}}{e^{s_1} + e^{s_2}}. \tag{3}$$

When a pair of adjectives is used, the ambiguity of one prompt is reduced by its antonym, casting the task as a binary classification where the final score is regarded as a relative similarity [24]. Although this method effectively leverages the prior knowledge of CLIP, the predicted quality score is solely dependent on the contrastive similarity, which is not accurate. Therefore, in our design, we improve the network by using a regression model $R(\cdot)$ to predict the quality score, enhancing the precision of the prediction and better adapting CLIP to the quality assessment task after further reducing ambiguity with more fine-grained quality-related adjectives.

3.2 Overview of *CLIP-AGIQA*

The overall framework of our method is shown in Fig. 2. *CLIP-AGIQA* consists of four components: learnable context, quality category, image quality regression, and the text encoder and image encoder in CLIP. In addition to the regression design, to better utilize the prior knowledge of the CLIP model, we incorporate learnable context for fine-tuning, inspired by the CoOp approach [37]. We also introduce additional quality category to address the ambiguity issues mentioned in CLIP-IQA. These two types of text-related information together form supplementary textual information to assist CLIP in adapting to the task of generative image quality assessment.

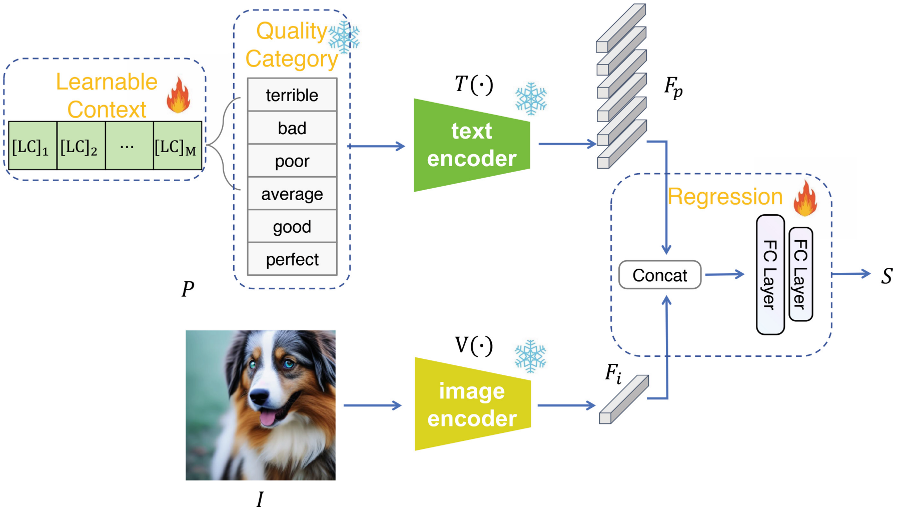


Fig. 2. Overall Architecture of *CLIP-AGIQA*.

Learnable Context. Since prompt engineering is a significant challenge in the application of CLIP, and the design of prompts can greatly impact performance, even with extensive manual tuning, the resulting prompts are by no means guaranteed to be optimal for downstream tasks [37]. Therefore, we abandon traditional subjective prompt settings in favor of a learnable prompt strategy. CLIP is sensitive to the choice of prompts, so we need to design a suitable set to leverage its prior knowledge. Similar to CoOp [37], we avoid manual prompt adjustments by modeling the context words using continuous vectors, which are end-to-end learned from the data, while freezing a large number of CLIP's pre-trained parameters. Specifically, as shown in Fig. 2, we use learnable context. We employ a unified context version from CoOp, where all prompts share the same context. The prompt design for the text encoder $T(\cdot)$ is as follows:

$$P = [LC]_1[LC]_2 \dots [LC]_M[QC] \tag{4}$$

Each $[LC]_m$ $(m \in \{1, \dots, M\})$ is the learnable context, represented as a vector with the same dimensionality as the word embeddings (i.e., 512 for CLIP). Here, M is a hyperparameter specifying the number of context tokens.

Text Encoder and Image Encoder. We utilize the text encoder $T(\cdot)$ and image encoder $V(\cdot)$ from CLIP. The text encoder is based on a Transformer architecture [22] and is responsible for generating text representations from natural language. In contrast, the image encoder is designed to map high-dimensional images into a low-dimensional embedding space. This encoder's architecture can resemble a CNN like ResNet-50 [5] or a Vision Transformer (ViT) [1]. In our setup, we employ these encoders separately to process our input textual information P and image information I, generating intermediate features used to predict quality score.

Quality Category. Due to the inherent language ambiguity in quality assessment tasks, utilizing CLIP as a versatile prior for visual perceptual evaluation is not straightforward. Similar to the antonym design in CLIP-IQA, we employ a series of quality-related auxiliary categories in Eq. (4) $[QC]$ to enhance the expression of the quality assessment task by describing the goodness of quality in a finer granularity. When using a set of quality-related adjective categories, they align with the correct category akin to the antonym prompts in CLIP-IQA, thereby reducing ambiguity. This transforms the task into multi-class classification, where the final score can be regarded as relative similarity, calculated through regression rather than using softmax as in CLIP-IQA. Specifically, we utilize six adjectives-terrible, bad, poor, average, good, and perfect-as quality category words to reduce ambiguity, thus better leveraging CLIP's priors. In addition, we also explore in the Sect. 4.3 the impact of the number and types of different words on its effectiveness. This design, together with the setting of the first learnable context, constitutes additional textual information to assist CLIP in transferring to the task of generated image quality assessment.

Image Quality Regression. To better fit the CLIP features to the data distribution for the task of evaluating the quality scores of generated images, we follow the paradigm of general quality assessment tasks by using the regression model $R(\cdot)$ to predict quality scores. We concatenate the image features $F_i = V(I) \in \mathbb{R}^{1\times N}$ and the textual features $F_p = T(P) \in \mathbb{R}^{6\times N}$ as the input features F.

$$F = concat(F_i, F_p) \tag{5}$$

We then process the concatenated features F through two fully connected (FC) layers. Here, the parameters of the FC layers are also learnable. The projection sizes are from 7 * 512 to 512 and from 512 to 1, respectively. Finally, we obtain the predicted quality score S, expressed as follows:

$$S = R(F) \tag{6}$$

Throughout the entire learning process, we employ the Mean Squared Error (MSE) as the loss function, with the specific formula shown below:

$$L = \frac{1}{N}\sum_{i=1}^{n}(S - y)^2 \tag{7}$$

where S represents the predicted quality score, and y represents the ground truth of the quality score.

4 Experiments

4.1 Experimental Settings

Datasets. To validate the effectiveness of our method, we conduct evaluations on two quality assessment benchmarks for generated images: AGIQA-3K [14] ans AIGCIQA2023 [25]. AGIQA-3K is a database containing 2,982 AI-generated images produced by six different models, including GAN-based, auto-regression-based, and diffusion-based models and subjective experiments are organized to obtain MOS (Mean Opinion Score) labels in terms of perceptual quality, which range from 0 to 5. AIGCIQA2023 collects over 2000 images using 100 prompts and six state-of-the-art text-to-image generation models, and quality and authenticity ratings are obtained by subjective experiments, which are ultimately scaled to a range of 0–100.

Evaluation Metrics. We use three common metrics in image quality assessment: PLCC, SRCC, and KRCC. PLCC (Pearson Linear Correlation Coefficient) measures the linear relationship between the predicted quality scores and the subjective scores. SRCC (Spearman Rank Correlation Coefficient) measures the consistency in the ranking order between the predicted quality scores and the subjective scores. KRCC (Kendall Rank Correlation Coefficient) measures the consistency in pairwise comparisons between the predicted quality scores and the subjective scores. All three metrics range from $[-1, 1]$, with values closer to 1 indicating higher correlation.

Training Details. The proposed *CLIP-AGIQA* is implemented in PyTorch and trained on 1 NVIDIA A100 GPU. ViT-B/16 [1] is used as the image encoder's backbone, and SGD is applied to optimize the network with an initial learning rate of 0.002. The training process was conducted over 100 epochs with a batch size of 32 and a learnable context length of 16. For learning rate scheduling, we employed a cosine annealing strategy, allowing the learning rate to decrease gradually throughout the training. Additionally, we implemented a warm-up phase during the first epoch, where the learning rate was held constant at 1×10^{-5}.

4.2 Experiment on Different Datasets

We focus on exploring the potential of *CLIP-AGIQA* in overall quality perception assessment. We conduct experiments on two widely used AGIQA benchmarks: AGIQA-3K [14] ans AIGCIQA2023 [25]. We also compare *CLIP-AGIQA* with different IQA methods, including handcrafted-based methods such as CEIQ [32], NIQE [17] and BRISQUE [16], and several learning-based methods like DBCNN [35], CLIP-IQA [24] and CNNIQA [9].

Table 1 presents the performance results of different IQA models on AGIQA-3K database, demonstrating that *CLIP-AGIQA* shows strong performance. As

Table 1. Comparison with the state-of-the-art IQA methods on AGIQA-3K dataset. The best performance results are marked in RED and the second-best performance results are marked in BLUE

Methods	PLCC	SRCC	KRCC
FID [7]	0.1860	0.1733	0.1158
CEIQ [32]	0.4166	0.3228	0.2220
NIQE [17]	0.5171	0.5623	0.3876
GMLF [31]	0.8181	0.6987	0.5119
CNNIQA [9]	0.8469	0.7478	0.5580
DBCNN [35]	0.8759	0.8207	0.6336
CLIP-IQA [24]	0.8053	0.8426	0.6468
CLIPAGIQA(Ours)	0.8978	0.8618	0.6776

we can see, *CLIP-AGIQA* achieves PLCC, SRCC, KRCC values of 0.8978, 0.8618 and 0.6776, respectively. These results outperform all compared methods, showcasing the great potential of our approach.

Table 2 shows the comparison between our *CLIP-AGIQA* and other IQA methods on the AIGCIQA2023 dataset. It can be seen that our method not only meets or exceeds state-of-the-art performance in evaluating the quality of generated images but also significantly outperforms other IQA models in assessing the authenticity of the dataset, which refers to the ability to evaluate whether an image is AI-generated. This indicates that our model excels not only in quality assessment but also has great potential to extend to other aspects of evaluating generated images.

Figure 3 shows that *CLIP-AGIQA* is able to assess overall perceptual quality to a level comparable to human judgment. It can assign reasonable scores based on the quality of the generated images. Notably, this model demonstrates several interesting capabilities. For instance, in the first column of the first row, where a strange bowl appears in the scenery image, it identifies common flaws in generated images and assigns a low score. Similarly, although the person in the second column of the second row looks lifelike, the model may detect subtle defects such as issues with the fingers and assigns a relatively low score. The first and second column of the third row also receive a low score maybe due to unrealistic elements and detail issues.

4.3 Ablation Studies

As described in Sect. 3.2, we make three unique modifications to adapt CLIP for the quality assessment task. In this section, to verify the effectiveness of the proposed key components, we train five variants of *CLIP-AGIQA* in AGIQA-3K:

I) Without regression and using cosine similarity instead (following CoOp and using classification loss for tuning the context); II) Changing the backbone

Table 2. Comparison with the state-of-the-art IQA methods on AIGCIQA2023 dataset. The best performance results are marked in RED and the second-best performance results are marked in BLUE

Methods	Quality			Authenticity		
	PLCC	SRCC	KRCC	PLCC	SRCC	KRCC
NIQE [17]	0.5218	0.5060	0.3420	3954	3715	2453
BRISQUE [16]	0.6389	0.6239	0.4291	0.4796	0.4705	0.3142
HOSA [30]	0.6561	0.6317	0.4311	0.4985	0.4716	0.3101
CNNIQA [9]	0.7937	0.7160	0.4955	0.5734	0.5958	0.4085
Resnet18 [5]	0.7763	0.7583	0.5360	0.6528	0.6701	0.4740
VGG16 [20]	0.7973	0.7961	0.5843	0.6807	0.6660	0.4813
VGG19 [20]	0.8402	0.7733	0.5376	0.6565	0.6674	0.4843
CLIPAGIQA(Ours)	0.8302	0.8140	0.5991	0.7797	0.7940	0.5849

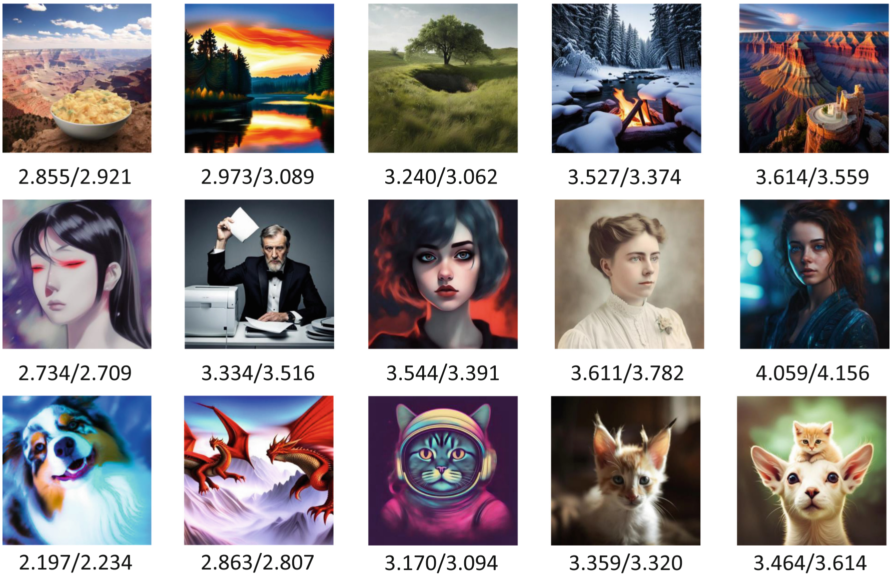


Fig. 3. *CLIP-AGIQA* for assessing overall perceptual quality. Left: Model Scores, Right: Human Scores

network; III) Changing the length of learnable contexts; IV) Changing the length of quality categories; V) Changing the type of quality categories.

The results indicate that removing or changing any single factor leads to a decrease in performance, confirming their contribution to the performance results in Table 3. It is worth noting that CLIP-IQA$^+$ [24] has already validated the importance of learnable context and quality categories, so we only test the impact of regression on CLIP in the quality assessment of generated images. In

Table 3. Ablation Study Results

No.	Ablation	Setting	PLCC	SRCC	KRCC
0	full model	ViT-B/16, 16, 6 adjectives	0.8978	0.8618	0.6776
1	without regression	ViT-B/16, 16, 6 adjectives	0.8183	0.8201	0.6693
2	- (backbone)	ViT-B/32, 16, 6 adjectives	0.8954	0.8614	0.6751
		ResNet-101, 16, 6 adjectives	0.8837	0.8544	0.6665
3	- (context length)	ViT-B/16, 8, adjective	0.8951	0.8595	0.6746
		ViT-B/16, 32, 6 adjectives	0.8962	0.8605	0.6751
4	- (category length)	ViT-B/16, 16, 8 adjectives	0.8962	0.8616	0.6766
5	- (category type)	ViT-B/16, 16, 6 scores	0.8958	0.8604	0.6747

variant 1, we observed a significant improvement when regression is added. This indicates that the combination of CLIP priors with a simple regression model is already effective.

In variants 2–5, although the impact on the model's performance is minimal, exploring these variants still provides us with valuable insights to understand and improve *CLIP-AGIQA*. Variants 2 and 3 are set up similarly to those explored in CoOp [37]. In our investigation of the backbone, we find a similar conclusion: the more advanced the backbone, the better the performance. However, the conclusion from CoOp that having more context tokens leads to better performance is not satisfied when the context length increased from 16 to 32. This can be due to the increased number of parameters making it harder for the model to converge to an appropriate state, warranting further investigation in future work. Additionally, we demonstrate that a "good" initialization does not make much difference, though this is not explicitly included in the table.

In variants 4 and 5, when the length of quality categories increases indefinitely, the task intuitively becomes a one-to-one classification task, yet the performance does not improve. Possible reasons could be that having too many quality categories makes synonyms indistinguishable, or the model parameters are insufficient to differentiate between categories. Changing the type of quality categories to numbers representing score relationships results in a performance drop, likely because CLIP rarely uses numbers in training, making it difficult to directly represent score magnitudes with numbers.

5 Conclusion

In this paper, we propose *CLIP-AGIQA*, a model that effectively adapts to new assessment requirements for generated images by leveraging CLIP's comprehensive visual and textual knowledge. Directly using CLIP has limitations and does not align well with the task of generated image quality assessment. To address this, we design various categories representing different quality levels to input into CLIP's text encoder, mitigating semantic ambiguities. By introducing a learnable prompts strategy and utilizing multiple quality-related auxiliary categories, we fully exploit CLIP's textual knowledge. Our regression network directly maps CLIP features to quality scores, effectively combining CLIP's capabilities with the task of generated image quality assessment, thereby enhancing the model's performance. Experiments demonstrate that *CLIP-AGIQA*, when trained with different datasets, performs excellently in both datasets. Ablation studies confirm the effectiveness of the proposed components. In the future, we will further improve our work by developing CLIP's own weights during training or by using multiple learnable contexts to explore multi-dimensional, fine-grained quality scores.

Acknowledgments. This work is supported by the National Natural Science Foundation of China (NSFC) under Grant No. 62272460, Beijing Natural Science Foundation under Grant No. 4232037.

References

1. Dosovitskiy, A., et al.: An image is worth 16×16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
2. Frolov, S., Hinz, T., Raue, F., Hees, J., Dengel, A.: Adversarial text-to-image synthesis: a review. Neural Netw. **144**, 187–209 (2021)
3. Gu, S., Bao, J., Chen, D., Wen, F.: Giqa: generated image quality assessment. arXiv preprint arXiv:2003.08932 (2020)
4. Gu, X., Lin, T.Y., Kuo, W., Cui, Y.: Open-vocabulary object detection via vision and language knowledge distillation. arXiv preprint arXiv:2104.13921 (2021)
5. 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)
6. 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)
7. 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)
8. Hou, J., et al.: Towards transparent deep image aesthetics assessment with tag-based content descriptors. IEEE Trans. Image Process. (2023)
9. Kang, L., Ye, P., Li, Y., Doermann, D.: Convolutional neural networks for no-reference image quality assessment. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1733–1740 (2014)
10. Ke, J., Ye, K., Yu, J., Wu, Y., Milanfar, P., Yang, F.: Vila: learning image aesthetics from user comments with vision-language pretraining. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 10041–10051 (2023)
11. Khurana, D., Koli, A., Khatter, K., Singh, S.: Natural language processing: state of the art, current trends and challenges. Multimed. Tools Appl. **82**(3), 3713–3744 (2023)
12. Kirstain, Y., Polyak, A., Singer, U., Matiana, S., Penna, J., Levy, O.: Pick-a-pic: an open dataset of user preferences for text-to-image generation. Adv. Neural. Inf. Process. Syst. **36**, 36652–36663 (2023)

13. Li, B., Weinberger, K.Q., Belongie, S., Koltun, V., Ranftl, R.: Language-driven semantic segmentation. arXiv preprint arXiv:2201.03546 (2022)
14. Li, C., et al.: Agiqa-3k: an open database for ai-generated image quality assessment. IEEE Trans. Circuits Syst. Video Technol. (2023)
15. Liu, X., Van De Weijer, J., Bagdanov, A.D.: Rankiqa: learning from rankings for no-reference image quality assessment. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 1040–1049 (2017)
16. Mittal, A., Moorthy, A.K., Bovik, A.C.: No-reference image quality assessment in the spatial domain. IEEE Trans. Image Process. **21**(12), 4695–4708 (2012)
17. Mittal, A., Soundararajan, R., Bovik, A.C.: Making a "completely blind" image quality analyzer. IEEE Signal Process. Lett. **20**(3), 209–212 (2012)
18. Radford, A., et al.: Learning transferable visual models from natural language supervision. In: International Conference on Machine Learning, pp. 8748–8763. PMLR (2021)
19. Salimans, T., Goodfellow, I., Zaremba, W., Cheung, V., Radford, A., Chen, X.: Improved techniques for training gans. Advances in neural information processing systems **29** (2016)
20. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)
21. Soundararajan, R., Bovik, A.C.: Rred indices: reduced reference entropic differencing for image quality assessment. IEEE Trans. Image Process. **21**(2), 517–526 (2011)
22. Vaswani, A., et al.: Attention is all you need. Advances in neural information processing systems **30** (2017)
23. Vinker, Y., Pajouheshgar, E., Bo, J.Y., Bachmann, R.C., Bermano, A.H., Cohen-Or, D., Zamir, A., Shamir, A.: Clipasso: semantically-aware object sketching. ACM Trans. Graph. (TOG) **41**(4), 1–11 (2022)
24. Wang, J., Chan, K.C., Loy, C.C.: Exploring clip for assessing the look and feel of images. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 37, pp. 2555–2563 (2023)
25. Wang, J., Duan, H., Liu, J., Chen, S., Min, X., Zhai, G.: Aigciqa2023: A large-scale image quality assessment database for ai generated images: from the perspectives of quality, authenticity and correspondence. In: CAAI International Conference on Artificial Intelligence, pp. 46–57. Springer (2023)
26. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: from error visibility to structural similarity. IEEE Trans. Image Process. **13**(4), 600–612 (2004)
27. Wu, J., Lin, W., Shi, G., Li, L., Fang, Y.: Orientation selectivity based visual pattern for reduced-reference image quality assessment. Inf. Sci. **351**, 18–29 (2016)
28. Wu, X., Sun, K., Zhu, F., Zhao, R., Li, H.: Human preference score: Better aligning text-to-image models with human preference. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 2096–2105 (2023)
29. Xu, J., e al.: Imagereward: learning and evaluating human preferences for text-to-image generation. Advances in Neural Information Processing Systems **36** (2024)
30. Xu, J., Ye, P., Li, Q., Du, H., Liu, Y., Doermann, D.: Blind image quality assessment based on high order statistics aggregation. IEEE Trans. Image Process. **25**(9), 4444–4457 (2016)
31. Xue, W., Mou, X., Zhang, L., Bovik, A.C., Feng, X.: Blind image quality assessment using joint statistics of gradient magnitude and laplacian features. IEEE Trans. Image Process. **23**(11), 4850–4862 (2014)

32. Yan, J., Li, J., Fu, X.: No-reference quality assessment of contrast-distorted images using contrast enhancement. arXiv preprint arXiv:1904.08879 (2019)
33. Yuan, J., Cao, X., Li, C., Yang, F., Lin, J., Cao, X.: Pku-i2iqa: an image-to-image quality assessment database for ai generated images. arXiv preprint arXiv:2311.15556 (2023)
34. Zhang, C., Zhang, C., Zhang, M., Kweon, I.S.: Text-to-image diffusion models in generative ai: a survey. arXiv preprint arXiv:2303.07909 (2023)
35. Zhang, W., Ma, K., Yan, J., Deng, D., Wang, Z.: Blind image quality assessment using a deep bilinear convolutional neural network. IEEE Trans. Circuits Syst. Video Technol. **30**(1), 36–47 (2018)
36. Zhang, W., Zhai, G., Wei, Y., Yang, X., Ma, K.: Blind image quality assessment via vision-language correspondence: a multitask learning perspective. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 14071–14081 (2023)
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)

CoDeiT: Contrastive Data-Efficient Transformers for Deepfake Detection

John Zakkam[1], Umarani Jayaraman[1](✉), Subin Sahayam[2], and Ajita Rattani[3]

[1] Indian Institute of Information Technology Design and Manufacturing
Kancheepuram, Chennai, India
{ced18i059,umarani}@iiitdm.ac.in

[2] Shiv Nadar University, Chennai, India
subinsahayamm@snuchennai.edu.in

[3] University of North Texas, Denton, TX, USA
ajita.rattani@unt.edu

Abstract. Deepfakes, synthetic media manipulated using AI, pose significant challenges to credibility and security. With the increasing sophistication of deepfake generation, robust detection methods are crucial. In this paper, CoDeiT (Contrastive Data-efficient Transformers) is introduced, a framework for deepfake detection integrating a hierarchical attention mechanism in HiLo Transformer architecture with contrastive learning. It uses HiLo Attention to separate high-frequency (Hi-Fi) and low-frequency (Lo-Fi) information, enhancing computational efficiency and detection accuracy. The contrastive learning framework further increases discriminative power by maximizing the similarity between genuine instances and minimizing it between genuine and fake ones. Extensive data augmentation improves robustness across diverse datasets. Comprehensive experiments on benchmark datasets validate CoDeiT's effectiveness. Three variations of the architecture have been proposed: CoDeiT-S, CoDeiT-L, and CoDeiT-XL, each differing in the number of parameters and attention heads. CoDeiT-XL has achieved 86.9% accuracy and 0.95 AUC on DFDC, and 78.5% accuracy and 0.89 AUC on the challenging CelebDF dataset when trained on the FaceForensic++ dataset. It outperformed all state-of-the-art deepfake detection methods. CoDeiT is effective for deepfake detection due to its unique architecture and ability to capture both high-frequency and low-frequency information efficiently. The combination of high and low-frequency information allows the CoDeiT to extract rich and detailed features from the data. This dual focus is particularly effective in detecting subtle inconsistencies and manipulations present in deepfakes.

Keywords: Deepfake · HiLo Transformer · Contrastive Learning · CelebDF Dataset · Face Forensics++ · DFDC

A. Antonacopoulos et al. (Eds.): ICPR 2024, LNCS 15332, pp. 62–77, 2025.
https://doi.org/10.1007/978-3-031-78125-4_5

1 Introduction

A deepfake is a synthetic media creation, typically in the form of a video or audio recording, generated using artificial intelligence (AI) techniques, particularly deep learning. The term “deepfake” is a combination of “deep learning” and “fake.” These AI-generated media are designed to convincingly mimic real people, often making it appear as though they are saying or doing something they never actually did. There has been a recent increase in videos, often obscene, where faces are swapped with others using neural networks, known as deepfakes[1], which have become a significant public concern[2]. The availability of open-source software and apps for face swapping has resulted in a large number of synthetically generated deepfake videos surfacing on social media and in the news, creating a major technical challenge for their detection and filtering. Consequently, the creation of effective tools to automatically detect these videos with swapped faces is of utmost importance.

1.1 Need for Deepfake Detection

The need for deepfake detection is critical due to the significant risks associated with their misuse. Deepfakes can spread misinformation and disinformation, manipulate political outcomes, and incite public panic or unrest. They pose severe threats to privacy and reputation, such as in cases of revenge porn and defamation, and undermine trust in digital media by making it difficult to discern authentic content. The detection of deepfakes is crucial to counteract their potential to mislead, harm, and erode trust in digital media. It is essential for protecting individuals’ privacy and reputation, maintaining public trust, ensuring national security, and upholding legal and ethical standards. Developing and deploying robust deepfake detection technologies is vital to mitigate these risks and safeguard society against the malicious use of this powerful technology.

Consequently, the identification of deepfakes has attracted a lot of attention in recent years. Recent developments in deep learning [19] have made it more challenging for humans to identify deepfakes. Current deepfake detection methods face significant limitations, including poor generalization across different datasets, vulnerability to adversarial attacks, high computational costs, and a lack of interpretability. These challenges hinder their reliability and practical deployment. Segregating high and low-frequency information is crucial because high-frequency details capture subtle artifacts and fine-grained textures indicative of manipulations, while low-frequency information provides the broader contextual integrity of the image or video. This balance between detailed local analysis and comprehensive global understanding enhances the model’s ability to detect deepfakes accurately and robustly, improving performance and generalization across diverse scenarios.

[1] Open source: https://github.com/deepfakes/faceswap.

[2] BBC report (Feb 3, 2018): http://www.bbc.com/news/technology-42912529.

1.2 Contributions

In this paper, an approach for deepfake detection has been proposed. The proposed approach is based on the HiLo Transformer architecture with a contrastive learning framework. The effectiveness of the proposed approach has been demonstrated through a series of experiments on multiple benchmark datasets. The main contributions of the work are:

- Introduction of the HiLo Transformer Architecture for Efficient Deepfake Detection: The HiLo Transformer's design focuses on high-frequency details locally and low-frequency structures globally, reducing redundant processing and enhancing data efficiency. Its hierarchical attention mechanism ensures effective feature extraction from each image or video frame.
- Integration of a Contrastive Learning Framework to Enhance Model Discriminative Power: Contrastive learning boosts the HiLo transformer's ability to distinguish real from fake images by learning robust features from contrasting pairs. Data augmentation enhances generalization with limited labeled data.
- Comprehensive Evaluation on Multiple Benchmark Datasets Demonstrating State-of-the-Art Performance: Evaluations on benchmark datasets show the HiLo transformer's superior performance, setting new state-of-the-art results. This combination improves accuracy and data efficiency, effective with smaller training datasets.

The rest of the paper is organized as follows. The related work is discussed in Sect. 2. Next, the proposed architecture is discussed in Sect. 3. The experimental results are discussed in Sect. 4. At last, conclusions and future work are discussed in Sect. 5.

2 Related Work

Initial attempts [11,15] used a combination of CNNs and LSTMs to learn temporal patterns of extracted features. The work [15] uses CNN-based methods to detect the difference in the resolution between warped faces and its surroundings. There have also been attempts to analyze the Photo Response Non-Uniformity (PRNU) noise patterns in forged images in work [14]. Another approach is performing mesoscopic level analysis in [1]. However, with the increasing quality of forged content, the performance of these detectors becomes challenging.

Over the years, CNN-based methods have become popular. These methods focus on comparing features contrasting near the blending boundary. Although these CNN-based methods perform well, CNN with LSTM methods are competent.

Recent methods still use CNNs to accurately detect deepfakes leveraging the concept of attention. Approaches [8] have also used transformers in this context to distinguish the identities of forged images near the blending boundary. In [12] self-supervised methods to detect deepfakes works are currently performing better compared to supervised models in the cross-dataset testing setting.

Contrastive learning has emerged as a powerful paradigm for representation learning, particularly in the context of unsupervised and self-supervised learning. The fundamental idea behind contrastive learning is to bring similar instances closer in the representation space while pushing dissimilar instances apart. This approach has been successfully applied in various domains, including computer vision and natural language processing.

The InfoNCE loss, introduced by Oord et al. [20], is a popular objective function in contrastive learning that aims to maximize the mutual information between different views of the same data. Chen et al. [4] further advanced this concept with SimCLR, demonstrating the effectiveness of simple yet powerful augmentations and a contrastive loss for visual representation learning. MoCo (Momentum Contrast) by He et al. [13] introduced a dynamic dictionary with a queue and a moving-averaged encoder, significantly improving the scalability of contrastive learning methods.

In the context of deepfake detection, contrastive learning frameworks have been leveraged to enhance the discriminative power of models by learning robust feature representations that differentiate genuine content from manipulated media. Works such as Li et al. [3] have shown the efficacy of self-supervised learning in improving the generalization of deepfake detection models across diverse datasets. Grill et al. [10] introduced BYOL (Bootstrap Your Own Latent), which avoids the use of negative pairs and demonstrates state-of-the-art performance in self-supervised learning. Caron et al. [2] presented SwAV (Swapping Assignments between Views), which combines contrastive learning with clustering to improve feature learning. Zbontar et al. [28] proposed Barlow Twins, which uses redundancy reduction to achieve competitive performance without requiring large batch sizes or negative pairs.

HiLo Attention, proposed by Pan et al. [21], extends the capabilities of vision transformers by disentangling high and low frequency patterns, enhancing the efficiency and effectiveness of attention mechanisms. This approach has been particularly beneficial for tasks requiring detailed analysis of fine-grained features and broader contextual understanding.

In the realm of deepfake detection, the application of vision transformers offers a promising direction, leveraging their ability to model intricate visual patterns and capture subtle inconsistencies in manipulated media. Works such as ViT-G by Kolesnikov et al. [9] and DeIT by Touvron et al. [26] have shown that vision transformers can achieve competitive performance with efficient training strategies.

3 Proposed Architecture: Extending HiLo Transformer with Contrastive Learning

Extending the HiLo Transformer with contrastive learning can enhance its ability to distinguish between authentic and fake data by learning more discriminative feature representations. Contrastive learning is a self-supervised learning technique that aims to maximize the similarity between related data points (positive pairs) while minimizing the similarity between unrelated data points (negative pairs). This section explains enhancing performance by integrating contrastive learning with the HiLo Transformer. The flow diagram of the proposed architecture is shown in Fig. 1. It has several different stages which are explained below.

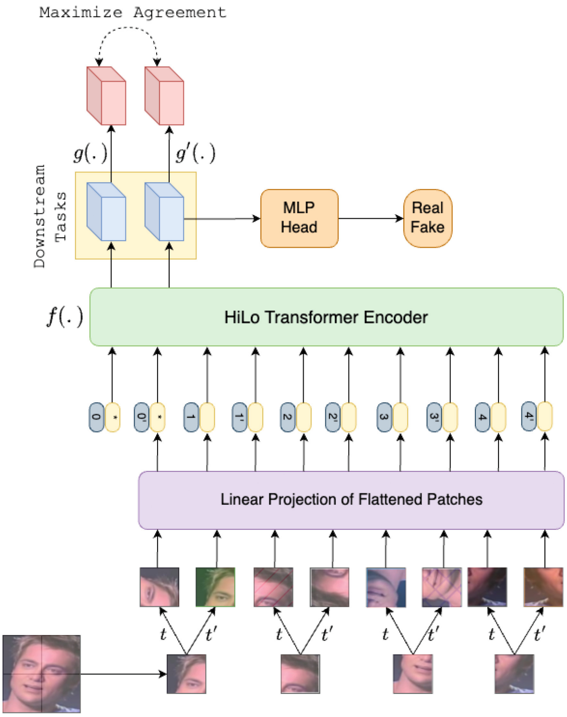


Fig. 1. Proposed CoDeiT architecture.

3.1 Data Preparation

The dataset is obtained from publicly available sources namely Celeb-DF [16], FaceForensics++ (FF++) [23], and Deepfake Detection Challenge (DFDC) [7]. All these datasets contain both real and deep fake videos. Frames are extracted at regular intervals to create images. It ensures the dataset has diverse examples of different types of deep fakes.

3.2 HiLo Transformer Architecture

The traditional transformer is replaced with the HiLo Transformer with a hierarchical attention mechanism namely local attention and global attention.

- Local Attention (high-frequency): The model processes small patches or regions of the input to capture fine-grained details.
- Global Attention (low-frequency): The model aggregates information from the entire input to maintain contextual understanding.

HiLo Attention is motivated by the observation that natural images contain rich frequencies where high/low frequencies play different roles in encoding image patterns, (i.e.) local fine details and global structures, respectively [21]. A typical multi-head self-attention (MSA) layer enforces the same global attention across all image patches without considering the characteristics of different underlying frequencies. HiLo Attention separates an MSA layer into two paths: High-frequency attention (Hi-Fi) and Low-frequency attention (Lo-Fi) as shown in Fig. 2.

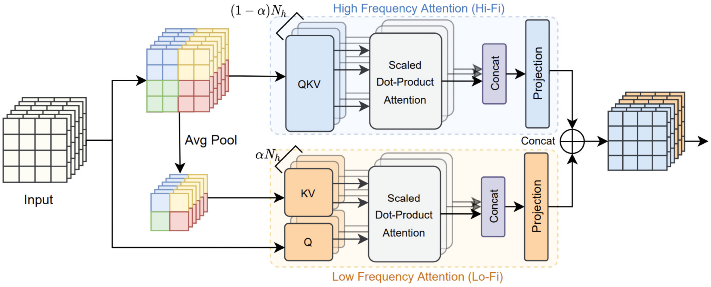


Fig. 2. Framework of HiLo attention. N_h refers to the total number of self-attention heads at this layer. α denotes the split ratio for high/low frequency heads; reproduced from [22].

Multi-head Self-Attention (MSA): The MSA mechanism operates by splitting the input into multiple heads, each learning different representations. The attention calculation is defined as:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{1}$$

where Q (queries), K (keys), and V (values) are linearly transformed versions of the input, and d_k is the dimensionality of the keys.

HiLo Attention: HiLo Attention extends MSA by disentangling high and low frequency patterns. The process involves:

Linear Transformations: First, apply linear transformations to the input to get the query $\mathbf{Q}$, key $\mathbf{K}$, and value $\mathbf{V}$ matrices:

$$\mathbf{Q} = \mathbf{X}\mathbf{W}_Q, \quad \mathbf{K} = \mathbf{X}\mathbf{W}_K, \quad \mathbf{V} = \mathbf{X}\mathbf{W}_V \tag{2}$$

where $\mathbf{W}_Q, \mathbf{W}_K, \mathbf{W}_V \in \mathbb{R}^{d\times d}$ are learnable weight matrices.

Step 2: Compute Hi-Fi Attention. Hi-Fi Attention focuses on local details. This can be done using standard scaled dot-product attention:

$$\text{Attention}_{HiFi}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d}}\right)\mathbf{V} \tag{3}$$

Step 3: Compute Lo-Fi Attention. Lo-Fi Attention focuses on global structures. One way to implement this is by using a different set of linear transformations that aggregate information over larger regions:

$$\mathbf{Q}_{\text{Lo}} = \mathbf{X}\mathbf{W}_{Q,\text{Lo}}, \mathbf{K}_{\text{Lo}} = \mathbf{X}\mathbf{W}_{K,\text{Lo}}, \mathbf{V}_{\text{Lo}} = \mathbf{X}\mathbf{W}_{V,\text{Lo}} \tag{4}$$

$$\text{Attention}_{LoFi}(\mathbf{Q}_{\text{Lo}}, \mathbf{K}_{\text{Lo}}, \mathbf{V}_{\text{Lo}}) = \text{softmax}\left(\frac{\mathbf{Q}_{\text{Lo}}\mathbf{K}_{\text{Lo}}^T}{\sqrt{d}}\right)\mathbf{V}_{\text{Lo}} \tag{5}$$

Step 4: Combine Hi-Fi and Lo-Fi Attention. Finally, combine the outputs of Hi-Fi and Lo-Fi attention:

$$\begin{aligned}\text{HiLo Attention}(\mathbf{X}) = {} & \mathbf{W}_{\text{HiFi}}\text{Attention}_{HiFi}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) \\ & + \mathbf{W}_{LoFi}\text{Attention}_{\text{LoFi}}(\mathbf{Q}_{\text{Lo}}, \mathbf{K}_{\text{Lo}}, \mathbf{V}_{\text{Lo}})\end{aligned} \tag{6}$$

where $\mathbf{W}_{\text{HiFi}}$ and $\mathbf{W}_{\text{LoFi}}$ are learnable weights that balance the contributions of Hi-Fi and Lo-Fi attention. The advantages of using HiLo Transformer for deep fake detection are as follows.

- Enhanced Sensitivity to Artifacts: The HiLo Transformer's local attention mechanism can detect fine-grained artifacts commonly present in deep fakes, such as subtle pixel-level anomalies and inconsistencies in facial features.
- Improved Contextual Analysis: The global attention mechanism allows the model to understand and analyze the entire image or frame context, making it capable of identifying broader inconsistencies, such as unnatural expressions or movements.
- Scalability and Efficiency: By efficiently handling high-resolution images and video frames, the HiLo Transformer can process and analyze large datasets, which is essential for robust deep fake detection.
- Flexibility: The architecture's adaptability enables it to handle various types of deep fakes, including different styles and techniques used to generate fake images and videos.

Three variations of the architecture have been proposed: CoDeiT-S, CoDeiT-L, and CoDeiT-XL, each differing in the number of parameters and attention heads. The configurations of these architectures are given in Table 1.

These variations have been introduced to assess the performance of the proposed architecture at different computational costs. The larger models, CoDeiT-L and CoDeiT-XL, offer better accuracy but demand more computational resources. CoDeiT-S balances performance and efficiency, making it suitable for resource-constrained environments.

Table 1. Comparison of model variations in terms of parameters, FLOPs, and attention heads.

Model	Params	FLOPs	Attention Heads
CoDeiT-S	22M	4.8B	6
CoDeiT-L	86M	15.1B	12
CoDeiT-XL	307M	60.9B	24

3.3 Contrastive Learning Framework

- Feature Extraction: The HiLo Transformer is used to extract features from the input data. The hierarchical attention mechanism ensures that both local details and the global context are captured.
- Projection Head: A projection head is added on top of the HiLo Transformer to map the extracted features to a lower-dimensional space suitable for contrastive learning.

To further enhance detection capability, a contrastive learning framework is integrated with the HiLo Transformer. Contrastive learning has demonstrated remarkable success in unsupervised and self-supervised learning tasks by maximizing the similarity between positive pairs and minimizing it between negative pairs [4,13]. In this work, a contrastive loss function, specifically InfoNCE (Information Noise Contrastive Estimation) loss, is defined as follows:

$$L_{\text{InfoNCE}} = -\log \frac{\exp(z_i \cdot z_j/\tau)}{\sum_{k=1}^{2N} \mathbf{1}_{[k\neq i]} \exp(z_i \cdot z_k/\tau)} \tag{7}$$

where z_i and z_j are the encoded representations of a positive pair, τ is a temperature parameter that scales the logits, and $\mathbf{1}_{[k\neq i]}$ is an indicator function equal to 1 if $k \neq i$ and 0 otherwise. The InfoNCE loss maximizes the agreement between positive pairs while minimizing the agreement with negative pairs, thereby learning a feature space where similar instances are closely clustered, and dissimilar instances are further apart.

In this framework, the latent space learns representations by aligning similar instances and pushing apart dissimilar ones. This is particularly useful for deepfake detection, as it enhances the discriminative capability of the model by learning subtle differences between genuine and fake instances.

The advantages of using the HiLo Transformer with contrastive learning for deep fake detection are as follows.

- Enhanced Feature Discrimination: Contrastive learning helps the HiLo Transformer learn more discriminative features, improving its ability to distinguish between similar and dissimilar instances, such as real and fake data.
- Robustness to Variations: By learning to differentiate between positive and negative pairs, the model becomes more robust to variations and noise in the data, leading to better generalization.

- Improved Data Efficiency: Contrastive learning leverages unlabeled data effectively, reducing the reliance on large labeled datasets and improving performance even with limited labeled data.

3.4 Training Process

- Pretraining with Contrastive Learning: Train the HiLo Transformer with the contrastive loss on a large dataset to learn robust feature representations.
- Fine-Tuning: After pretraining, fine-tune the model on a labeled dataset for the specific task, such as deep fake detection, using a cross-entropy loss which is a supervised loss function.

3.5 Overall Framework

The overall deepfake detection framework involves the following steps:

- The HiLo Transformer is used to extract hierarchical features from input images or videos. These features capture various levels of abstraction, making it easier to identify subtle inconsistencies characteristic of deepfakes.
- The extracted features are projected into an embedding space suitable for contrastive learning. In this space, the contrastive loss function operates, helping the model learn discriminative representations that are effective for deepfake detection.
- The model is trained using a dataset containing both genuine and deepfake instances. Its performance is evaluated across different datasets to ensure robustness and generalizability.
- Robust data augmentation techniques, such as random cropping, flipping, color jittering, and noise addition, create diverse training samples. These augmentations expose the model to a wide variety of data, enhancing its ability to generalize across different datasets and improving its robustness in real-world applications.

By combining the hierarchical design and HiLo Attention of the HiLo Transformer with a robust contrastive learning framework, the proposed method achieves high precision and efficiency in deepfake detection tasks.

4 Experimental Results

4.1 Datasets

We evaluated our method on three widely used deepfake detection datasets: Celeb-DF [16], FaceForensics++ (FF++) [23], and Deepfake Detection Challenge (DFDC) [7]. The Celeb-DF dataset consists of 5,639 high-quality deepfake videos, split into 70% for training and 30% for testing. The FaceForensics++ dataset comprises manipulated videos generated using various methods, divided into a 70% training set and a 30% testing set. The DFDC dataset is a large-scale collection of 100,000 videos, with 60% for training, 20% for validation, and 20% for testing as described in 2.

Table 2. Deepfake Datasets with Train-Test Split

Dataset	# of Images	Train Split	Test Split	Real: DF
Celeb-DF	2,342,200	1,639,540	702,660	590: 5,639
(FF++)	1,019,800	713,860	305,940	1,000: 1,000
DFDC	2,271,700	1,363,020	454,340	1,131: 4,113

4.2 Training Setup

The codebase is built on the PyTorch framework using the timm deep learning library. All experiments have been conducted on a Linux machine with a 40GB NVIDIA A100 GPU. The networks are trained using Cross Entropy Loss and optimized with SGD with momentum, an initial learning rate of 0.005, momentum of 0.9, and a mini-batch size of 512. The larger batch size was selected to efficiently utilize the available GPU RAM.

4.3 Evaluation Metrics

To evaluate the performance of the proposed approach, accuracy, and AUC are used. Accuracy is the ratio of correctly predicted instances to the total instances, providing a basic measure of correctness. However, the primary metric we focus on is the Area Under the Receiver Operating Characteristic Curve (AUC-ROC). AUC-ROC is preferred as it provides an aggregate measure of performance across all classification thresholds, offering a comprehensive view of the model's ability to distinguish between genuine and fake instances. Unlike accuracy, which can be misleading in imbalanced datasets, AUC-ROC considers the true positive rate and false positive rate, making it a more reliable metric for comparing the performance of deepfake detection models. By focusing on AUC-ROC, the evaluation captures the nuanced performance of the model in identifying deepfakes, crucial for real-world applications.

4.4 Cross-dataset Evaluation

To demonstrate the effectiveness of the proposed approach in learning robust latent representations, it is performed cross-dataset evaluations. The proposed architecture is trained on the FF++ dataset and evaluated on the Celeb-DF and DFDC datasets, and vice versa. The results show that CoDeiT architecture, particularly CoDeiT-XL, has achieved the highest accuracy and AUC-ROC scores across all evaluations. This indicates that the proposed architecture can generalize well across different datasets.

Trained on FF++ and Tested on Others: The first set of experiments involved evaluating video-level deep fake detection accuracy and AUC of CoDeiT architecture (CoDeiT-S, CoDeiT-L, and CoDeiT-XL). These architectures are trained on FF++ and tested on DFDC and CelebDF datasets using high-quality videos and the results are given in Table 3. The existing method results are taken from their references. It can be observed that all three versions of CoDeiT architectures outperformed state-of-the-art methods, including MesoNet [1], Xception [5], Efficient-B7 [25], FFD [17], ISPL [18], Seferbekovv [24], ResNet + LSTM [29], and Efficient-B1 + LSTM [27]. Among the three, the CoDeiT-XL architecture has achieved the highest accuracy and AUC of 86.9 and 0.95 for DFDC (HQ) dataset while it is 78.5 and 0.89 for the CelebDF (HQ) dataset respectively. It demonstrates **superior performance** in detecting deepfakes compared to other methods.

Table 3. Trained on FF++ and Tested on DFDC and CelebDF HQ videos: Comparing performance (Accuracy/AUC) of the proposed architecture with existing state-of-the-art methods

Model	DFDC (HQ)		CelebDF (HQ)	
	Accuracy	AUC	Accuracy	AUC
MesoNet [1]	53.6	0.74	50.1	0.75
Xception [5]	72.0	0.79	77.2	0.88
Efficient-B7 [25]	71.8	0.78	71.4	0.80
FFD [17]	63.1	0.69	69.2	0.76
ISPL [18]	69.6	0.78	71.2	0.83
Seferbekov [24]	72.0	0.85	75.3	0.86
ResNet + LSTM [29]	61.2	0.67	58.2	0.72
Eff.B1 + LSTM [27]	67.2	0.75	75.3	0.84
ID-Reveal [6]	80.4	0.91	71.6	0.84
CoDeiT-S	**82.5**	**0.92**	**73.8**	**0.85**
CoDeiT-L	**84.7**	**0.94**	**76.1**	**0.87**
CoDeiT-XL	**86.9**	**0.95**	**78.5**	**0.89**

Trained on DFDC and Tested on Others: The second experiment evaluates the performance of the proposed architecture trained on DFDC and tested on the FF++ and CelebDF datasets with high-quality videos and the results are given in Table 4. The CoDeiT architecture demonstrated superior performance compared to other methods across all three versions. Among the three, the CoDeiT-XL architecture has achieved the highest accuracy and AUC of 88.1 and 0.95 for FF++ (HQ) dataset while it is 78.6 and 0.87 for the CelebDF (HQ) dataset respectively. This demonstrates the robustness of CoDeiT when trained on different datasets, showcasing its ability to learn effective representations for deepfake detection.

Trained on Celeb-DF and Tested on Others: The third experiment evaluates the performance of the proposed architecture trained on Celeb-DF and tested on the FF++ and DFDC datasets with high-quality videos and, the results are given in Table 5. The CoDeiT architecture demonstrated superior performance compared to other methods across all three versions. Among the three, the CoDeiT-XL architecture has achieved the highest accuracy and AUC of 85.6 and 0.93 for FF++ (HQ) dataset while it is 86.1 and 0.95 for the DFDC (HQ) dataset respectively. This demonstrates the robustness of CoDeiT when trained on different datasets, showcasing its ability to learn effective representations for deepfake detection. Figure 3 shows the accuracy graph obtained for the cross-dataset evaluation of CoDeiT with the top four state-of-the-art methods. This evaluation was conducted across two datasets DFDC (HQ) and CelebDF (HQ).

Table 4. Trained on DFDC and tested on FF++ and CelebDF HQ videos: Comparing performance (Accuracy/AUC) of the proposed architecture with existing state-of-the-art methods

Model	FF++ (HQ)		CelebDF (HQ)	
	Accuracy	AUC	Accuracy	AUC
MesoNet [1]	55.4	0.58	50.1	0.75
Xception [5]	74.1	0.81	77.2	0.88
Efficient-B7 [25]	72.6	0.80	71.4	0.80
FFD [17]	64.3	0.71	69.2	0.76
ISPL [18]	70.8	0.79	71.2	0.83
Seferbekov [24]	73.5	0.86	75.3	0.86
ResNet + LSTM [29]	62.4	0.70	58.2	0.72
Eff.B1 + LSTM [27]	68.3	0.77	75.3	0.84
ID-Reveal [6]	81.7	0.92	71.6	0.84
CoDeiT-S	**83.4**	**0.89**	**73.9**	**0.82**
CoDeiT-L	**85.8**	**0.93**	**76.2**	**0.84**
CoDeiT-XL	**88.1**	**0.95**	**78.6**	**0.87**

4.5 Ablation Study

To justify the model architecture choices, an extensive ablation study has been conducted. Variations in the number of attention heads, MLP layers, and other hyperparameters are explored to evaluate their impact on the model's performance. These are summarized in Table 6. The ablation study demonstrates that the chosen architecture and hyperparameters effectively balance performance and computational efficiency. Increasing the number of attention heads improves the model's ability to capture complex patterns and subtle inconsistencies in

Table 5. Comparing performance (Accuracy/AUC) of our models with existing state-of-the-art models on FF++ and DFDC HQ videos, trained on Celeb-DF.

Model	FF++ (HQ)		DFDC (HQ)	
	Accuracy	AUC	Accuracy	AUC
MesoNet [1]	55.4	0.58	60.1	0.72
Xception [5]	55.6	0.58	77.2	0.88
Efficient-B7 [25]	54.9	0.59	71.4	0.80
FFD [17]	54.4	0.56	69.2	0.76
ISPL [18]	56.6	0.59	71.2	0.83
Seferbekov [24]	58.3	0.62	75.3	0.86
ResNet + LSTM [29]	55.0	0.58	65.2	0.78
Eff.B1 + LSTM [27]	57.2	0.62	75.3	0.84
ID-Reveal [6]	78.3	0.87	79.6	0.90
CoDeiT-S	**80.4**	**0.89**	**81.2**	**0.91**
CoDeiT-L	**83.2**	**0.91**	**83.7**	**0.93**
CoDeiT-XL	**85.6**	**0.93**	**86.1**	**0.95**

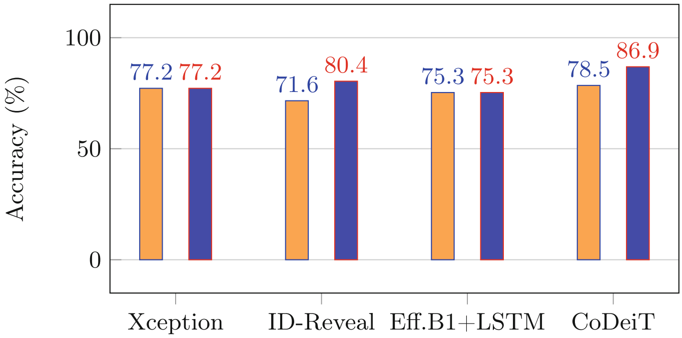


Fig. 3. CoDeiT-XL: Accuracy of cross-dataset is shown trained on FaceForensic++ tested across two datasets with SOTA Methods. The orange bars represent DFDC (HQ) and the blue bars represent CelebDF (HQ). (Color figure online)

deepfake content, but this also increases the computational cost. Similarly, more MLP layers allow the model to learn more complex representations, but the performance gains diminish beyond 4 layers, indicating an optimal balance between model complexity and performance. The larger models, CoDeiT-L and CoDeiT-XL, provide better accuracy at the cost of increased computational cost, making them more suitable for offline analysis. CoDeiT-S, with 22 M parameters, provides a good balance between efficiency and performance, making it suitable for real-time applications.

Table 6. Ablation study results showing the impact of different configurations of attention heads and MLP layers on CoDeiT-XL trained on FF++. Best results are indicated in bold.

Attention Heads	MLP Layers	Train Acc. (%)
6	2	85.3
6	4	87.1
12	2	88.4
12	4	89.7
24	4	91.2
24	**8**	**91.5**

5 Conclusions

In this paper, CoDeiT, a novel framework for deepfake detection, is introduced that leverages the strengths of the hierarchical attention mechanism and contrastive learning. The proposed Hierarchical Data-efficient Transformer (HiLo Transformer) employs HiLo Attention to effectively disentangle and process high and low-frequency information, significantly enhancing the model's ability to detect subtle manipulations indicative of deepfakes. A contrastive learning framework using the InfoNCE loss function was incorporated, which further improved the discriminative power of the model by maximizing the similarity between genuine instances and minimizing the similarity between genuine and fake instances. The use of comprehensive data augmentation techniques ensured robustness and generalizability across diverse datasets. Three variations of the architecture have been proposed: CoDeiT-S, CoDeiT-L, and CoDeiT-XL, each differing in the number of parameters and attention heads. CoDeiT-XL has achieved 86.9% accuracy and 0.95 AUC on DFDC, and 78.5% accuracy and 0.89 AUC on the challenging CelebDF dataset when trained on the FaceForensic++ dataset. Extensive experiments on widely used deepfake detection datasets, including Celeb-DF, FaceForensics++, and the Deepfake Detection Challenge, demonstrated that CoDeiT outperforms existing state-of-the-art methods across all three versions.

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. Caron, M., Misra, I., Mairal, J., Goyal, P., Bojanowski, P., Joulin, A.: Unsupervised learning of visual features by contrasting cluster assignments. In: NeurIPS (2020)
3. Chen, L., Zhang, Y., Song, Y., Liu, L., Wang, J.: Self-supervised learning of adversarial example: Towards good generalizations for deepfake detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 18710–18719 (2022)

4. Chen, T., Kornblith, S., Norouzi, M., Hinton, G.: A simple framework for contrastive learning of visual representations. In: International Conference on Machine Learning, pp. 1597–1607. PMLR (2020)
5. Chollet, F.: Xception: Deep learning with depthwise separable convolutions. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2017)
6. Cozzolino, D., Rössler, A., Thies, J., Nießner, M., Verdoliva, L.: Id-reveal: identity-aware deepfake video detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 15108–15117 (2021)
7. Dolhansky, B., et al.: The deepfake detection challenge (dfdc) dataset. arXiv preprint arXiv:2006.07397 (2020)
8. Dong, X., et al.: Protecting celebrities from deepfake with identity consistency transformer. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 9468–9478 (2022)
9. Dosovitskiy, A., et al.: An image is worth 16×16 words: transformers for image recognition at scale. In: International Conference on Learning Representations (2020)
10. Grill, J.B., et al.: Bootstrap your own latent: a new approach to self-supervised learning. In: NeurIPS (2020)
11. Güera, D., Delp, E.J.: Deepfake video detection using recurrent neural networks. In: 2018 15th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS), pp. 1–6. IEEE (2018)
12. Haliassos, A., Mira, R., Petridis, S., Pantic, M.: Leveraging real talking faces via self-supervision for robust forgery detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 14950–14962 (2022)
13. He, K., Fan, H., Wu, Y., Xie, S., Girshick, R.: Momentum contrast for unsupervised visual representation learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 9729–9738 (2020)
14. Koopman, M., Rodriguez, A.M., Geradts, Z.: Detection of deepfake video manipulation. In: The 20th Irish Machine Vision and Image Processing Conference (IMVIP), pp. 133–136 (2018)
15. Li, Y., Lyu, S.: Exposing deepfake videos by detecting face warping artifacts. arXiv preprint arXiv:1811.00656 (2018)
16. Li, Y., Yang, X., Sun, P., Qi, H., Lyu, S.: Celeb-df: a large-scale challenging dataset for deepfake forensics. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 3207–3216 (2020)
17. Nguyen, D.T., Yamagishi, J., Echizen, I.: Ffd: faceforensics dataset. arXiv preprint arXiv:1911.08854 (2019)
18. Nguyen, H.N., Zhou, L.A., Nguyen, H.H.: Ispl: improved synthesis for personal lip movement. In: Proceedings of the 25th ACM International Conference on Multimedia (2017)
19. Nguyen, T.T., et al.: Deep learning for deepfakes creation and detection: a survey. Comput. Vis. Image Underst. **223**, 103525 (2022)
20. Oord, A.v.d., Li, Y., Vinyals, O.: Representation learning with contrastive predictive coding. arXiv preprint arXiv:1807.03748 (2018)
21. Pan, Z., Cai, J., Zhuang, B.: Fast vision transformers with hilo attention. Adv. Neural. Inf. Process. Syst. **35**, 14541–14554 (2022)
22. Pan, Z., Cai, J., Zhuang, B.: Fast vision transformers with hilo attention (2023). https://arxiv.org/abs/2205.13213

23. Rössler, 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)
24. Seferbekov, S.: Seferbekov: deepfake detection challenge. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) (2020)
25. Tan, M., Le, Q.V.: Efficientnet: rethinking model scaling for convolutional neural networks. In: International Conference on Machine Learning, pp. 6105–6114. PMLR (2019)
26. Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A., Jégou, H.: Training data-efficient image transformers & distillation through attention. In: International Conference on Machine Learning, pp. 10347–10357 (2021)
27. Wang, W., Zhang, Z., Zhang, J.: Eff.b1 + lstm: efficientnet and long short-term memory for video forgery detection. In: Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW) (2019)
28. Zbontar, J., Jing, L., Misra, I., LeCun, Y., Deny, S.: Barlow twins: self-supervised learning via redundancy reduction. In: ICML (2021)
29. Zhao, H., Ge, S., Li, Y.: Resnet + lstm: combining residual networks with long short-term memory for face forgery detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2019)

A Lightweight and High-Fidelity Model for Generalized Audio-Driven 3D Talking Face Synthesis

Shunce Liu[1,2], Yuwei Zhong[3], Huixuan Wang[4(✉)], and Jingliang Peng[1,2(✉)]

[1] Shandong Provincial Key Laboratory of Network Based Intelligent Computing,
University of Jinan, Jinan, China
ise_pengjl@ujn.edu.cn

[2] School of Information Science and Engineering, University of Jinan, Jinan, China

[3] School of Software and Microelectronics, Peking University, Beijing, China

[4] School of Software, Shandong University, Jinan, China
hx.wang@sdu.edu.cn

Abstract. We present LW-GeneFace, a lightweight and high-fidelity model for generalized audio-driven facial animation, in this paper. We develop this model by reducing the size while maintaining the synthetic quality of GeneFace, an audio-driven facial animation model known for its high fidelity and generalization capabilities. Specifically, we compress the first and the third stages of GeneFace as they dominate the model size. In the first stage, we propose a lightweight version of the WaveNet-based network inspired by MobileNetV3 and DP-block. It utilizes depthwise separable convolution and dual-path feature extraction to compress the network while maintaining effective feature extraction. The shared network structure in the dual-path feature extraction further reduces model complexity and improves training efficiency.

In the third stage, we generate realistic 3D renderings at reduced model size by introducing novelties in RAD-NeRF. Technically, we reduce the hash table sizes in the grid-based encoding modules, as well as present a lightweight bottleneck MLP architecture to increase the non-linearity of the model. Experimental results demonstrate that LW-GeneFace achieves state-of-the-art performance with both model size and synthetic quality considered. The source code of LW-GeneFace will be released after acceptance of this paper.

Keywords: audio-driven facial animation · talking face synthesis · lightweight WaveNet · lightweight RAD-NeRF · bottleneck MLP

1 Introduction

The intersection of Artificial Intelligence (AI) and content creation has led to a vibrant field known as Artificial Intelligence Generated Content (AIGC), which

S. Liu and Y. Zhong—Equivalent Contribution author.

A. Antonacopoulos et al. (Eds.): ICPR 2024, LNCS 15332, pp. 78–90, 2025.
https://doi.org/10.1007/978-3-031-78125-4_6

has recently witnessed a surge in interest, particularly in the realm of audio-driven facial animation [5]. This burgeoning area of research is poised to revolutionize the way we perceive and interact with digital characters and environments, offering a seamless integration of speech and lifelike facial expressions.

Our research endeavors to harness the nuances of speech signals to generate highly realistic facial animations, thereby enabling a voice-controlled animation experience. The state-of-the-art in this domain has demonstrated the potential to produce facial expressions that not only mirror the spoken content but also reflect the emotional and tonal subtleties of human speech. Despite the strides made, challenges remain, particularly regarding the computational complexity and resource requirements of advanced models. The need for efficient processing on low-configuration computing platforms further compounds these challenges.

This study aims to address these concerns by developing a lightweight model that maintains the fidelity of lifelike facial animations based on audio input. We refine the GeneFace model [21], leveraging its renowned accuracy and adaptability, through novel optimizations that streamline its core components without compromising on the quality of synthesis. Major contributions of this work are summarized as follows.

- In the first stage of variational motion generator, we re-design the WaveNet-based network by employing depthwise separable convolution and dual-path feature extraction, and further making a Siamese structure of the dual paths. This design substantially reduces the model size with almost negligible compromise on motion estimation accuracy.
- In the third stage of NeRF renderer, we make optimizations by reducing the hash table sizes in the grid-base encoding modules and using bottleneck MLPs for promoted modeling capability. These optimizations lead to substantial reduction in model size with quality of animation synthesis well maintained.
- We conduct extensive experiments to show that our proposed model achieves state-of-the-art performance when both model size and animation synthesis quality are considered.

2 Related Work

2.1 Speech-Driven Facial Animation

The task of speech-driven facial animation aims to reproduce arbitrary input speech audio from a specific person. It has received considerable attention of the computer vision community in recent years. Researchers in the early stage predominantly utilized methodologies including cross-modal retrieval technology [17] and hidden Markov models [15]. These approaches were designed to establish the mapping between auditory speech signals and facial animation datasets, thereby facilitating the production of animated sequences. Nonetheless, such technologies introduce more rigorous requirements for the deployment environment of the model and compel the need for manual annotation of visual phonemes, which can be a labor-intensive process.

The rapid advancement of deep learning in recent years has greatly accelerated the progress of speech-driven facial animation technology. Zhou *et al.* [25] proposed the MakeItTalk model, which extracted the content and speaker information from the auditory input. Leveraging this data, the model predicts facial landmarks that reflect the speaker's dynamic expressions. Prajwal *et al.* [13] proposed a rigorous evaluation benchmark for measuring lip synchronization in unconstrained videos. Zhou *et al.* [24] modularized audio-visual representations through the formulation of an implicit, low-dimensional pose code. Lu *et al.* [11] employed a three-stage network structure to extract facial action poses from speech features and generate facial animations. Wang *et al.* [19] developed a model for predicting head posture using a recurrent neural network based on motion perception. The model extracts the low-frequency overall motion pattern of the head from the speech signal. Fan *et al.* [3] proposed the FaceFormer, a model leveraging the Transformer architecture to generate a sequence of semantic 3D facial animations. They designed two attention mechanisms to learn the connection and sequence dependence between speech and vision, respectively. Furthermore, they employed periodic positional encoding for input representation, performed deep encoding of the audio signal, and adopted an autoregressive prediction approach. Fang *et al.* [4] presented FE-GAN, a facial animation generation algorithm utilizing a generative adversarial network (GAN) framework, which integrates dual auxiliary classifiers along with a pair of recognizers to enhance the animation's fidelity. Ye *et al.* [22] proposed a dynamic convolution kernel (DCK) strategy to enhance convolutional neural networks. The approach utilizes a fully convolutional network featuring Dynamic Convolutional Kernels (DCKs), which is capable of real-time selection between two modalities: speech and video. This method yields high-quality facial animation videos from the data source. Shen *et al.* [14] designed a latent diffusion model for visual attention mechanisms. Ye *et al.* [21] put forward GeneFace, a three-stage framework that uses 3D facial landmarks as intermediate variables.

The aforementioned deep learning based approaches primarily concentrate on audio-visual synchronization or generalization capabilities. However, the lightweight performance of the model is relatively an oversight. Indeed, there remains potential for further improvement in the complexity management of the state-of-the-art approaches.

2.2 Neural Radiance Field for Face Rendering

NeRF (Neural Radiance Field) is a novel approach in the field of computer graphics and 3D modeling, which has been applied in various research projects for creating speech-driven facial animation. Guo *et al.* [6] proposed AD-NeRF, an approach that integrates audio signal characteristics into a conditional implicit function, thereby generating a dynamic neural radiance field. Yao *et al.* [20] presented DFA-NeRF, which combines neural radiance fields and the audio information. The model considers lip movement characteristics and personal attributes as two independent parts of the NeRF condition, and predicts lip movements synchronized with the corresponding speech content. Liu *et al.* [10] proposed Semantic-aware Speaking Portrait NeRF, which designs two semantic-aware modules to handle the local facial semantics and the global head-torso relationship. Based on the recent advancements in Grid-based NeRF, Tang *et al.* [16] established a novel decomposition of the complex, high-dimensional conversational portrait representation into three more tractable, low-dimensional feature grids.

Though flexible modeling and high-fidelity rendering has been achieved, there remains significant room for optimizing the NeRF models in both model size and modeling capability in the context of audio-driven facial animation.

3 Methodology

In this work, we are motivated to construct a lightweight and high-fidelity model for generalized audio-driven facial animation. We choose GeneFace [21] as the base and introduce novelties to optimize it for reduced model size with maintained level of animation synthesis quality.

The GeneFace model [21] consists of three stages in sequence: **audio-to-motion** stage that generates facial landmark positions from the input audio, **motion domain adaption** stage that refines the predicted 3D landmarks from the multi-speaker domain into the target person domain, and **motion-to-image** stage that renders high-fidelity frames guided by the 3D landmarks using a NeRF-based renderer.

As part of the data preprocessing, GeneFace [21] utilizes the pretrained HuBERT model [9] to extract audio features that are used in the first stage. In general, there are alternative audio feature extractors with highly varied performance and complexity characteristics [12], and different audio-driven 3D taking face synthesis models [3,11,21,25] have employed different pretrained audio feature extractors. In this work, we focus on compressing the GeneFace model less the HuBERT feature extractor and leave better choice or simplification of audio feature extractor for our future investigation.

Excluding the HuBERT feature extractor, the total parameters of the first and third stages amount to 24.620M, accounting for 96.87% of the overall parameter count across the three stages. As detailed in the Ablation Study in Sect. 4.4, these first and third stages dominate the model size. Consequently, we have reduced the complexity of these two stages in our model design and have adopted the same second stage structure as utilized in GeneFace. As a result, we construct a lightweight audio-driven facial animation model, which we name lightweight GeneFace (LW-GeneFace). Details of the first and the third stages of LW-GeneFace are given in the following subsections.

3.1 Variational Motion Generator

The first stage utilizes a variational auto-encoder (VAE) to complete the audio-to-motion transform and is named variational motion generator. HuBERT features [9] of the input audio wave are extracted and used as input to the motion

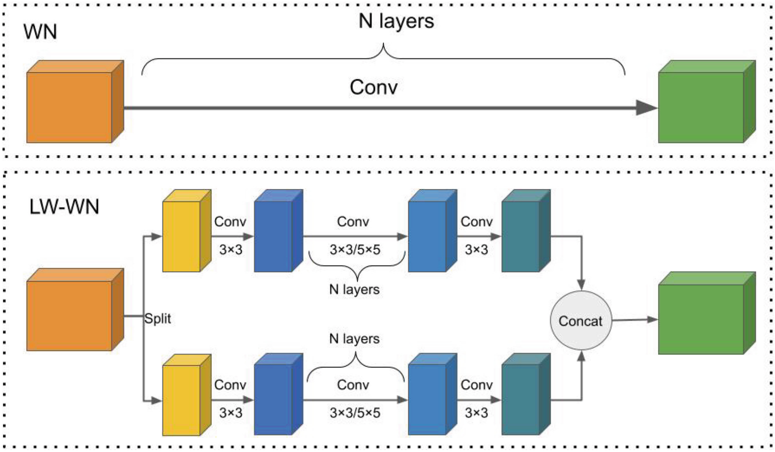


Fig. 1. Structures of the WN in GeneFace (top) and the LW-WN in LW-GeneFace (bottom).

generator, and the motion generator infers 3D positions of 68 facial landmarks at each frame that represent the facial motion.

Major components of the first stage in GeneFace include pretrained HuBERT feature extractor, flow-based prior, pretrained SyncNet, encoder and decoder. Due to the space limit, we refer the readers to the original paper [21] for detailed explanation. In our design of the first stage, we adopt the same framework but optimize the encoder and the decoder for reduced model size.

In the first stage of GeneFace, the encoder and decoder primarily use a structure similar to that of the WaveNet (WN) [18], as briefly shown in the top portion of Fig. 1. This module utilises multi-layer 5×5 and 3×3 convolutions and residual modules to extract speech features and predict facial landmarks positions, where the dilation factors of convolution incrementally increase with depth. In contrast to WN, our proposed lightweight WaveNet (LW-WN) model is inspired by MobileNetV3 [8] and DP-block [23] and made highly compact. It utilizes depthwise separable convolution and dual-path feature extraction to compress the WaveNet structure, as depicted in the bottom portion of Fig. 1 and explained below.

Drawing inspiration from MobileNetV3 [8], we modify WN by using more lightweight depthwise separable convolution in place of normal convolution. The multi-layer depthwise separable convolution utilises a combination of depth convolutions with N (N = 4 in the encoder and N = 8 in the decoder) layers of convolution kernels at size of 3×3 and 5×5 to enhance the learning of multi-scale features. Besides, we are inspired by DP-block [23] and develop a network structure for dual-path feature extraction. The feature tensors are divided equally and input into the two paths of the dual-path feature extraction network for further processing.

It is important to note that our dual-path feature extraction network utilises a Siamese structure. That is, the two paths share the same network structure and parameters. Regardless of which path the input data comes from, it will go through the same network structure for feature extraction. This design substantially reduces the model size and helps improve the training efficiency.

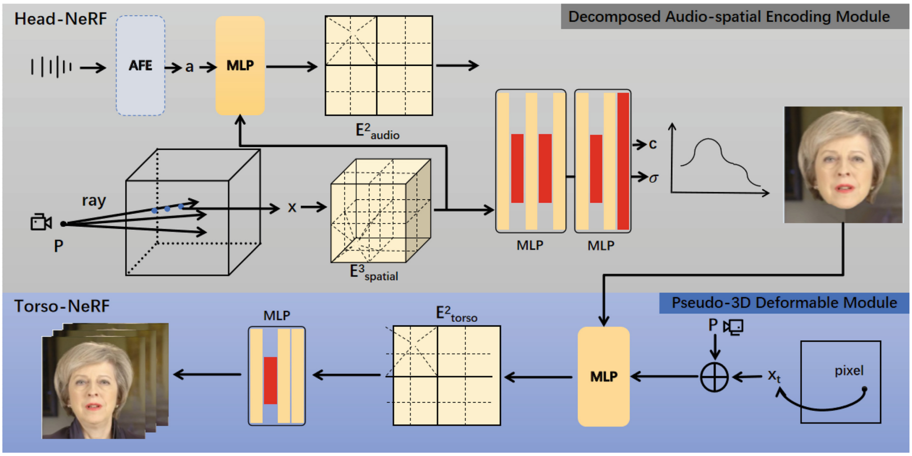


Fig. 2. Structure of the LW-RAD-NeRF model. AFE is an Audio Feature Extractor, $\mathbf{a}$ is an extracted audio feature, $\mathbf{x}$ is a 3D spatial coordinate, and $\mathbf{x}_t$ is a 2D torso coordinate. $E^3_{spatial}$, E^2_{audio} and E^2_{torso} are all gird encoders.The original MLP layers are colored yellow and the introduced ones are colored red. (Color figure online)

3.2 NeRF-Based Renderer

Following GeneFace, the proposed LW-GeneFace also employs RAD-NeRF [16] to render the head and the torso parts, respectively, in the third stage. The Head-NeRF is firstly trained and, thereafter, the torso-NeRF is trained with the rendering image of the Head-NeRF as background. Contrastively, we propose a lightweight NeRF-based renderer (LW-RAD-NeRF) with reduced model size while maintaining the quality of animation synthesis.

A key insight of LW-RAD-NeRF is to decompose the holisitc high-dimensional audio-guided protrait representation into separate low-dimensional trainable feature grids for simplified computation. Two NeRF modules, *i.e.* Decomposed Audio-spatial Encoding Module and Pseudo-3D Deformable Module, are designed to render the head and the torso, respectively. Both modules are grid-based NeRF models, where trainable features are associated with grid points and an arbitrary sample is encoded by the linear interpolation of the grid features. All the grid features are learned as network parameters. As shown in Fig. 2, two grid encoders, $E^3_{spatial}$ and E^2_{audio}, are used in the Decomposed

Audio-spatial Encoding Module to encode the 3D spatial coordinate $\mathbf{x}$ and the 2D audio coordinate $\mathbf{x}_a$, respectively, and one grid encoder, E^2_{torso} is used in the Pseudo-3D Deformable Module to encode one torso coordinate $\mathbf{x}_t$ per pixel. Note that, instead of a complete grid data structure, a hash table is used for each grid encoder to store the trainable feature vectors for memory efficiency, which is indexed by hashing the grid positions. Each trainable feature vector contains both color and density information, and increasing the hash table size should lead to more precise color and density modeling and better quality of rendering. Nevertheless, we observed that once the hash table size goes beyond a certain threshold, the model accuracy sees only marginal improvement while the model parameter count is significantly increased. As such, hash table size is one key factor that trades off representational quality and memory efficiency.

As model compactness is one of our primary goals and an overly large hash table has little impact on model accuracy, we propose to reduce the number of grid features at the first step. Specifically, we reduce the hash table sizes for all three grid encoders in the LW-RAD-NeRF models. Furthermore, we observe that there are complicated interactions among various portions of a face, and a high order of non-linearity should be involved for accurate modeling of facial animation. As such, we further propose an optimized MLP module for enhanced modeling capability, which we name bottleneck MLP. Technically, the bottleneck MLP module is constructed by integrating additional MLP layers into the two NeRF modules. These added layers include ones with diminished widths. The original MLP layers are denoted in yellow, while the newly introduced layers are indicated in red, as shown in Fig. 2. Note that, although we have increased the number of layers in the bottleneck MLP module, the computational increase is manageable due to the reduction in width of some layers.

To be specific, the original hash table sizes for $E^3_{spatial}$, E^2_{audio} and E^2_{torso} are all 2^{16}. We reduce them to 2^{10}, 2^{10} and 2^{12}, respectively. Further, we add MLP layers to form bottleneck MLPs, as shown in Fig. 2, for enhanced modeling capability at controlled computation increase. Note that the extra storage required by the introduced MLP layers is far less than that saved by the reduction of hash table size. As a net effect, we obtain a significantly reduced model size with quality of animation maintained.

4 Experimental Evaluation

4.1 Metrics and Datasets

We compare our LW-GeneFace with several leading approaches, including Wav2Lip [13], MakeItTalk [25], PC-AVS [24], LSP [11], AD-NeRF [6], and GeneFace [21]. To evaluate the precision of lip synchronization, we employ the landmark distance (LMD) [2], and the SyncNet confidence score (Sync) [13]. We utilize the Fréchet Inception Distance (FID) score [7] to measure the full image quality. To further assess the generalizability, a set of out-of-domain (OOD) audio tests are applied for all benchmark methods.

Additionally, we employ two metrics to evaluate the complexity of the model: the number of parameters (Param) and the computational cost in Giga Floating-Point Operations (GFLOPs) during inference. The number of parameters indicates the memory footprint for model storage. GFLOPs measures the model's computing load, reflecting the amount of floating-point calculations performed per second. In this experiment, we utilize an audio segment of approximately 10 s to test GFLOPs.

Regarding the dataset, we utilize the LRS3-TED corpus by Fouras *et al.* [1], to train the first two stages, *i.e.*, the variational motion generator and the post-net, of GeneFace and LW-GeneFace. Furthermore, a video featuring the target person speaking for several minutes with an accompanying audio track is required to facilitate the training of a NeRF-based person portrait renderer. In order to compare with the state-of-the-art baselines, we utilize the same five facial videos of GeneFace [21], each of which consists of a video with an average length of 6,000 frames, recorded at a frame rate of 25 frames per second (fps).

4.2 Implementation Details

The Adam optimizer was used during the training process with an initial learning rate of 1×10^{-4}, and β_1 and β_2 values of 0.9 and 0.999, respectively. The network was trained on one GPU (NVIDIA RTX 3090 24 GB) with 40k steps for the Variational Motion Generator, 20k steps for the postnet, and 500k iterations for the NeRF-based renderer.

4.3 Results and Analysis

We compare our method with the state-of-the-art audio-driven talking head animation baselines. All the test input audio sequences are unseen during training. We evaluate the quality of synthesized animations through quality metrics and a user study. Further, we evaluate the complexity of models through complexity metrics. Statics of these evaluations are provided in Tables 1, 2 and 3, which are analyzed in the following subsections, respectively.

Metrics-Based Quality of Synthesized Animation. The image synthesis quality of various algorithms is shown in Table 1, where the data of Wav2Lip, MakeItTalk, PC-AVS, LSP, and AD-NeRF are all from the GeneFace paper [21]. Since complete inference models for all target faces are not released by the authors of GeneFace, we retrained the GeneFace models for all target characters and placed the test results in Table 1.

First we compare our approach with image-based generation baselines which generate a talking-head video from one or several reference images. Specially, we compare with Wav2Lip, MakeItTalk and PC-AVS. We have the following observations. (1) Wav2Lip, MakeItTalk, and PC-AVS perform poorly on the FID metric due to low image fidelity. (2) Both the Sync and Sync(OOD) scores of Wav2Lip outperform that of ground truth's. An expert lip-sync discriminator is trained by Wav2Lip to suit the lip generation task. However, it synthesizes just a lower face patch and blend it into the target frame without taking target expressions and head-poses into account. Different from their paradigm, our method directly renders both full head and the background. (3) We achieve the best scores on the FID and FID (OOD) metrics, meaning that we can generate high-fidelity images for arbitrary audio sources.

Then we compare our method with model-based generation baselines including LSP, AD-NeRF and GeneFace. Among the compared methods, LSP, GeneFace and LW-GeneFace use 3D facial landmarks as an intermediate representation. AD-NeRF, GeneFace and LW-GeneFace are all based on NeRF while LSP utilizes an image-to-image translation network to generate animated videos. The statistics of the compared models are listed in Table 1. It can be seen that, while LSP, AD-NeRF, GeneFace and LW-GeneFace achieve roughly comparable performance in terms of image quality and generalizability, LW-GeneFace outperforms the rest by a large margin on FID and FID(OOD).

Table 1. Comparison with state-of-the-art methods. Key: [Best, Second Best, Third Best].

Method	FID ↓	LMD ↓	Sync ↑	FID(OOD) ↓	Sync(OOD) ↑
Wav2Lip	71.40	3.988	9.212	68.05	9.645
MakeItTalk	57.96	4.848	4.981	53.33	4.933
PC-AVS	96.81	5.812	6.239	98.31	6.156
LSP	29.30	4.589	6.119	35.21	4.320
AD-NeRF	27.52	4.199	4.894	35.69	4.225
GeneFace	28.22	4.321	5.412	28.66	4.372
Ground Truth	0	0	8.233	N/A	N/A
LW-GeneFace	26.52	4.679	5.471	26.73	4.314

User Study on Quality of Synthesized Animation. For this study, we sampled 2 audio clips in English and 2 audio clips in Korean and used them to synthesize the animations for Obama2 by all the 7 methods compared in Table 1. We engaged 26 participants and asked them to rank the synthesized animations (rendered as videos) in each of 5 aspects, *i.e.* lip-sync accuracy, image quality, naturalness of lip movement, emotion expression, and audio-expression synchronization. For each audio clip, seven animations were generated by all the methods. These animations were ranked by each participant in each aspect, respectively, with the 1st and best one receiving 7 points, the 2nd best one receiving 6 points, and so on, down to the 7th and worst one receiving 1 point. Corresponding to each method and aspect, the statistics of all the 26 participants rankings on all the 4 synthesized animations were gathered and marked in Table 2.

We make several key observations from Table 2. Firstly, LSP, AD-NeRF, GeneFace and LW-GeneFace are all person-specific methods and all achieved excellent image qualities. Secondly, while GeneFace achieved outstanding results in emotion expression and audio-expression synchronization, LW-GeneFace further advanced the performance. This should be attributed to the enhanced nonlinearity of modeling by the extended MLP structure and, probably, the deeper feature extraction by the devised LW-WN as well. Thirdly, Wav2Lip performs the best in lip-sync accuracy and naturalness of lip movement as it is specifically designed for the particular task of lip simulation.

Table 2. User study with different methods. The error bars are 95% confidence interval. Key: [Best, Second Best, Third Best].

Methods	Wav2Lip	MakeItTalk	PC-AVS	LSP	AD-NeRF	GeneFace	LW-GeneFace
Lip-sync Accuracy	6.77 ± 0.16	2.15 ± 0.10	5.62 ± 0.13	4.96 ± 0.22	1.31 ± 0.12	3.12 ± 0.10	4.08 ± 0.21
Image Quality	2.88 ± 0.18	3.23 ± 0.40	1.88 ± 0.22	6.00 ± 0.51	4.35 ± 0.13	4.88 ± 0.48	5.00 ± 0.17
Naturalness of Lip Movement	6.77 ± 0.11	2.15 ± 0.16	1.08 ± 0.07	5.92 ± 0.16	4.04 ± 0.11	3.08 ± 0.11	4.96 ± 0.21
Emotion Expression	4.88 ± 0.23	2.15 ± 0.10	1.00 ± 0.00	4.38 ± 0.20	3.19 ± 0.12	5.92 ± 0.07	6.64 ± 0.18
Audio-Expression Sync	4.04 ± 0.22	2.12 ± 0.12	1.04 ± 0.08	5.08 ± 0.22	3.19 ± 0.12	5.92 ± 0.15	6.62 ± 0.18

Table 3. Comparison with state-of-the-art methods on Parameter count and GFLOPs. Key: [Best, Second Best, Third Best].

Method	Wav2Lip	MakeItTalk	PC-AVS	LSP	AD-NeRF	GeneFace	LW-GeneFace
Param(M)↓	36.298	76.603	152.244	83.535	1.736	25.416	12.548
GFLOPs ↓	1,536	74,971	4,478	99,968	15,959,832	29,214	26,247

Model Complexity. We compare the proposed LW-GeneFace with the state-of-the-art approaches in terms of model complexity, as shown in Table 3. Considering that audio feature extractors with highly varied complexity characteristics may be alternatively utilized for audio-driven facial animation, the statistics in Table 3 are measured with the exclusion of the audio feature extractors to guarantee a fair comparison.

Comparing with image-based generation baselines, *i.e.*, Wav2Lip, MakeItTalk and PC-AVS, the proposed LW-GeneFace has the smallest parameter count, though Wav2Lip and PC-AVS have smaller GFLOPs. Note again that Wav2Lip and PC-AVS produce worse image fidelity than LW-GeneFace, as analyzed in Sect. 4.3.

Then we compare LW-GeneFace with model-based generation baselines, *i.e.*, LSP, AD-NeRF and GeneFace. The parameter count and GFLOPs of LSP are multiple times higher than those of LW-GeneFace. Although AD-NeRF has the smallest parameter count, it entails the largest amount of computing as shown by its GFLOPs. Comparing with GeneFace, LW-GeneFace reduces the model size (in terms of Param.) sharply by more than a half, with GFLOPs also showing a decrease. This clearly demonstrates the effect of lightweighting optimization by LW-GeneFace.

Table 4. Ablation study results. The settings are described in Sect. 4.4. Best results are in bold.

Setting	FID ↓	LMD ↓	Sync ↑	FID(OOD) ↓	Sync(OOD) ↑	Param(M) ↓ (VMG)	Param(M) ↓ (NBR)	GFLOPs ↓
LW-GeneFace	**26.52**	4.679	**5.471**	**26.73**	4.314	**11.425**	**0.327**	**26,247**
GeneFace	28.22	**4.321**	5.412	28.66	4.372	20.431	4.189	29,214
w. LW-WN	29.07	4.703	4.953	29.71	**4.495**	**11.425**	4.189	29,213
w. LW-RAD-NeRF	26.54	4.685	5.387	27.24	4.475	20.431	**0.327**	26,246

4.4 Ablation Study

In this section, we perform ablation experiments to demonstrate the necessity of each component in LW-GeneFace. Excluding the HuBERT feature extractor, the total parameters of the GeneFace base model amount to 25.416M, of which the first stage Variational Motion Generator (VMG) and the third stage NeRF-Based Renderer (NBR) account for 20.431M and 4.189M parameters, respectively. Since the volume of parameters in the first and third stages dominates the overall size of the model, we focus on evaluating the necessity of the first and third sections for the integrity of the whole LW-GeneFace model. The experimental results are shown in Table 4. Param(VMG) and Param(NBR) in the table represent the parameters of the first and the third stages of LW-GeneFace, respectively. Note that the data regarding complexity in Table 4 have been processed with the exclusion of the HuBERT extraction module.

We test two settings in this experiment. In the setting w. LW-WN, it can be seen that the addition of the LW-WN module slightly reduces the quality of synthesized animation, and the number of parameters is lowered to almost half of the base model GeneFace. In our dual-path feature extraction, although the two branches share parameters, the calculations required for each branch still need to be performed, so the amount of computation has not decreased much.

In the setting w. LW-RAD-NeRF, we can see that the addition of the LW-RAD-NeRF module achieves better quality of synthesized animation on metrics Sync(OOD), FID and FID(OOD). Benefiting from our proposed bottleneck MLPs, the model possesses generalization and high-fidelity capabilities. Furthermore, the metric Param(NBR) is reduced to one thirteenth, compared with that of the base model GeneFace.

5 Conclusion

In this work, we have proposed LW-GeneFace, a lightweight model for generalized and high-fidelity audio-driven 3D talking face synthesis by optimizing GeneFace. To be specific, we compress the first and the third stages of GeneFace, which dominate the model size, while introducing extra layers to the third stage for enhanced modeling capability with controlled growth of computation. For the first stage, we employ depthwise separable convolution and Siamese dual-path feature extraction to simplify the model. For the third stage, we reduce the hash table sizes for the grid-based encoders and enhance the MLP portions by bottleneck MLP modules, which result in a compact NeRF model for head and torso rendering. The experimental results demonstrated that LW-GeneFace achieves state-of-the-art performance when both model size and quality of animation synthesis are considered.

The limitations of our work are two-fold. On the one hand, LW-GeneFace is an offline model in essence. If this model is used for online interaction, it is necessary to wait for the user to finish a segment of speech before creating the corresponding character animation. Therefore, there will be a time delay between the user's speech and the character animation. It should be noted that this is also a problem with all the other algorithms we compared (see Table 2). On the other hand, with an augmentation in the number of layers in the bottleneck MLPs, the inference time cost also increases. For instance, in one of our tests, LW-GeneFace reaches 2.26 frames per second (fps) in animation synthesis while GeneFace achieves 2.99 fps.

Acknowledgments. This work was supported by Key R&D Program of Shandong Province, China (Competitive Innovation Platform) (No. 2023CXPT003) and Shandong Provincial Natural Science Foundation, China (No. ZR2022MF294).

References

1. Afouras, T., Chung, J.S., Zisserman, A.: Lrs3-ted: a large-scale dataset for visual speech recognition. arXiv preprint arXiv:1809.00496 (2018)
2. Chen, L., Li, Z., Maddox, R.K., Duan, Z., Xu, C.: Lip movements generation at a glance. In: Proceedings of the European conference on computer vision (ECCV), pp. 520–535 (2018)
3. Fan, Y., Lin, Z., Saito, J., Wang, W., Komura, T.: Faceformer: speech-driven 3d facial animation with transformers. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 18770–18780 (2022)
4. Fang, Z., Liu, Z., Liu, T., Hung, C.C., Xiao, J., Feng, G.: Facial expression gan for voice-driven face generation. The Visual Computer, pp. 1–14 (2022)
5. Gowda, S.N., Pandey, D., Gowda, S.N.: From pixels to portraits: a comprehensive survey of talking head generation techniques and applications (2023). https://arxiv.org/abs/2308.16041
6. Guo, Y., Chen, K., Liang, S., Liu, Y.J., Bao, H., Zhang, J.: Ad-nerf: audio driven neural radiance fields for talking head synthesis. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 5784–5794 (2021)
7. 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)

8. Howard, A., et al.: Searching for mobilenetv3. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 1314–1324 (2019)
9. Hsu, W.N., Bolte, B., Tsai, Y.H.H., Lakhotia, K., Salakhutdinov, R., Mohamed, A.: Hubert: Self-supervised speech representation learning by masked prediction of hidden units. IEEE/ACM Trans. Audio Speech Lang. Process. **29**, 3451–3460 (2021)
10. Liu, X., Xu, Y., Wu, Q., Zhou, H., Wu, W., Zhou, B.: Semantic-aware implicit neural audio-driven video portrait generation. In: European Conference on Computer Vision, pp. 106–125. Springer (2022)
11. Lu, Y., Chai, J., Cao, X.: Live speech portraits: real-time photorealistic talking-head animation. ACM Trans. Graph. (TOG) **40**(6), 1–17 (2021)
12. Mohamed, A., et al.: Self-supervised speech representation learning: a review. IEEE J. Sel. Top. Signal Process. **16**(6), 1179–1210 (2022)
13. Prajwal, K., Mukhopadhyay, R., Namboodiri, V.P., Jawahar, C.: A lip sync expert is all you need for speech to lip generation in the wild. In: Proceedings of the 28th ACM International Conference on Multimedia, pp. 484–492 (2020)
14. Shen, S., Zhao, W., Meng, Z., Li, W., Zhu, Z., Zhou, J., Lu, J.: Difftalk: crafting diffusion models for generalized audio-driven portraits animation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 1982–1991 (2023)
15. Sheng, C., et al.: Deep learning for visual speech analysis: a survey. arXiv preprint arXiv:2205.10839 (2022)
16. Tang, J., et al.: Real-time neural radiance talking portrait synthesis via audio-spatial decomposition. arXiv preprint arXiv:2211.12368 (2022)
17. 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)
18. Van Den Oord, A., et al.: Wavenet: a generative model for raw audio. arXiv preprint arXiv:1609.03499 (2016)
19. Wang, S., Li, L., Ding, Y., Fan, C., Yu, X.: Audio2head: audio-driven one-shot talking-head generation with natural head motion. arXiv preprint arXiv:2107.09293 (2021)
20. Yao, S., Zhong, R., Yan, Y., Zhai, G., Yang, X.: Dfa-nerf: personalized talking head generation via disentangled face attributes neural rendering. arXiv preprint arXiv:2201.00791 (2022)
21. Ye, Z., Jiang, Z., Ren, Y., Liu, J., He, J., Zhao, Z.: Geneface: generalized and high-fidelity audio-driven 3d talking face synthesis. arXiv preprint arXiv:2301.13430 (2023)
22. Ye, Z., et al.: Audio-driven talking face video generation with dynamic convolution kernels. IEEE Trans. Multimed. (2022)
23. Zhao, L., Wang, L.: A new lightweight network based on mobilenetv3. KSII Trans. Internet Inf. Syst. **16**(1) (2022)
24. Zhou, H., Sun, Y., Wu, W., Loy, C.C., Wang, X., Liu, Z.: Pose-controllable talking face generation by implicitly modularized audio-visual representation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 4176–4186 (2021)
25. Zhou, Y., Han, X., Shechtman, E., Echevarria, J., Kalogerakis, E., Li, D.: Makelttalk: speaker-aware talking-head animation. ACM Trans. Graph. (TOG) **39**(6), 1–15 (2020)

Illuminating the Dark: Unpaired Retinex and FFT-Based Low-Light Image Enhancement

Zeyu Li[1,2](✉)

[1] Institute of Software, Chinese Academy of Sciences, Beijing, China
lizeyu2023@iscas.ac.cn
[2] University of Chinese Academy of Sciences, Beijing, China

Abstract. In the field of computer vision, enhancing low-light images is a significant challenge, primarily due to the reliance on high-quality paired low-light and high-light images in supervised learning methods, which are expensive to acquire. This paper presents an unpaired approach for low-light image enhancement, integrating a Joint Estimation Network and a Multi-Domain Feature Fusion Network. The Joint Estimation Network is trained exclusively with pairs of low-light images of the same scene, while the Multi-Domain Feature Fusion Network is trained solely with normal-light images that are not paired with the aforementioned low-light images. The Joint Estimation Network decomposes low-light images into components of illumination, reflectance, and noise. After enhancing the illumination, it passes these components, along with the reflectance, to the Multi-Domain Feature Fusion Network. The Multi-Domain Feature Fusion Network employs multi-scale encoder-decoder modules and frequency domain adjustments to enhance details and maintain global consistency. Our method addresses the issues of insufficient illumination and high noise in low-light images, improving visual quality without the need for paired images, thereby increasing the model's practicality in real-world applications.

Keywords: Unpaired Low-light image enhancement · Retinex Theory · FFT

1 Introduction

Enhancing images captured in low-light environments is a fundamental task in computer vision, aimed at improving both the clarity and quality of such images. The complexity of this task arises from challenges like inherent noise, low contrast, and color distortions typical of low-light scenarios. Traditional techniques, including histogram equalization and methods based on the Retinex theory, often struggle to provide satisfactory outcomes. These methods may introduce undesirable artifacts or fail to adequately address the varying conditions found in low-light environments [19,30].

In recent years, deep learning-based techniques have achieved significant advancements in enhancing low-light images. Supervised learning methods,

A. Antonacopoulos et al. (Eds.): ICPR 2024, LNCS 15332, pp. 91–105, 2025.
https://doi.org/10.1007/978-3-031-78125-4_7

which depend on paired datasets of low-light and normal-light images, have notably improved visual quality. For instance, the work by [6] and the Zero-DCE model [15] utilize deep neural networks to predict enhancement functions effectively. Despite their success, these approaches are limited by the necessity for paired training data, which is often challenging and expensive to collect [24].

To overcome this limitation, unsupervised and zero-shot learning methods have emerged as promising alternatives, eliminating the need for paired training data. Methods such as RetinexDIP [48] and EnlightenGAN [18] leverage generative models and domain adaptation techniques to enhance low-light images without ground truth references. Despite their effectiveness, these methods may suffer from instability and artifacts due to the lack of direct supervision.

This paper presents an innovative unpaired approach for low-light image enhancement, overcoming the drawbacks of existing methods. Our approach is built on two key components: the Joint Estimation Network and the Multi-Domain Feature Fusion Network. The Joint Estimation Network breaks down low-light images into their illuminance, reflectance, and noise elements, enabling focused enhancement of each aspect. The Multi-Domain Feature Fusion Network integrates multiscale encoder-decoder modules with frequency domain adjustments to improve both local detail and global coherence, drawing inspiration from recent developments in frequency-based image processing [23,32]. Overall, The primary contributions of this paper can be summarized as follows:

- We propose a groundbreaking unpaired method for low-light image enhancement, which ingeniously integrates image decomposition with multi-domain feature fusion. This methodology allows for substantial enhancements in visual quality without relying on paired reference images, thereby offering a versatile solution for real-world deployment.
- Our Joint Estimation Network is architected to deconstruct low-light images into their essential components-illuminance, reflectance, and noise. This strategic decomposition facilitates precise and targeted enhancement of each constituent element, culminating in superior image quality. The network's design is meticulously optimized to capture and process the intricate interdependencies among these components, resulting in more accurate and aesthetically refined enhancements.
- The Multi-Domain Feature Fusion Network synergizes spatial and frequency domain information, employing Fast Fourier Transform (FFT) to preserve both fine-grained image details and overall coherence. This dual-domain strategy markedly improves the fidelity and texture of images, demonstrating state-of-the-art performance on benchmark datasets.

2 Related Works

2.1 Unsupervised Low-Light Image Enhancement

Unsupervised low-light image enhancement has gained substantial attention due to its ability to operate without paired training data, thus overcoming the limitations inherent in supervised learning. The LIME algorithm, which integrates Retinex theory with illumination map estimation, effectively enhances low-light images [16]. RetinexNet utilizes deep learning to decompose images into reflectance and illumination components, marking a significant advance in the field [38]. Inspired by tools like Photoshop, ExCNet introduced the concept of learning an "S-curve" for image enhancement [45]. Zero-DCE (Zero-Reference Deep Curve Estimation) built on this by proposing a zero-reference deep learning framework that enhances low-light images without requiring paired or unpaired data [15]. A survey by Li et al. highlighted the importance of various unsupervised and zero-shot methods, underlining their critical role in low-light image and video enhancement [24]. Additionally, Huang et al. proposed a Fourier-based enhancement method, demonstrating the efficacy of frequency domain techniques [17]. EnlightenGAN uses a U-Net architecture for the generator and employs dual discriminators to capture both global and local information [18]. RUAS focuses on modeling the intrinsic underexposed structures of low-light images, while RRDNet decomposes images into illumination, reflection, and noise components to achieve superior denoising effects [28,52]. SCI developed a cascaded illumination learning process with weight sharing, enhancing the robustness and effectiveness of the enhancement process [29]. NeRCo introduced multimodality into low-light image enhancement, expanding the capabilities of existing methods [41]. PairLIE, based on Retinex theory, uses pairs of low-light images for training and has achieved competitive results [13]. Despite these advancements, many methods rely on redundant loss functions to ensure convergence, which introduces numerous priors and limits their generalization ability. These developments illustrate the ongoing evolution of unsupervised low-light image enhancement techniques, each contributing to more effective and efficient solutions for enhancing images captured under challenging lighting conditions.

2.2 FFT-Based Image Enhancement

The Fast Fourier Transform (FFT) is a pivotal technique in image processing, enabling efficient conversion between the spatial and frequency domains, which facilitates various advanced methods. FFT-based low-pass and high-pass filtering manipulate frequency components to suppress noise and enhance image edges [1]. Fourier Low-Light Image Enhancement (FourLLIE) utilizes frequency information to enhance structural details and contrast in low-light images. Additionally, FFT is integral to image compression and reconstruction. By reducing redundancy in the frequency domain, it allows effective image restoration via inverse FFT, as demonstrated by Huang et al. [17] and Cai et al. [4]. In image registration, FFT employs phase correlation techniques to determine translational shifts between images, which is crucial for precise alignment in medical imaging [6]. These applications highlight FFT's versatility and power in advancing image processing technologies. Xu et al. [40] introduced a Fourier-based data augmentation technique aimed at improving domain generalization. Fuoli et al. [14] employed Fourier losses to restore high-frequency details in image super-resolution, while Yu et al. [44] leveraged Fourier frequency information for image dehazing. Similarly, Zhou et al. [49] applied these methods to pan-sharpening. Additionally, Zhou et al. [50] developed a Fourier-based up-sampling approach that enhances various computer vision tasks in a plug-and-play manner. At the same time, Huang et al. [17] and other researchers created Fourier-based algorithms for low-light image enhancement. Despite some limitations, these approaches demonstrate the wide-ranging applicability of Fourier frequency information. However, FFT-based techniques are not without their drawbacks. A significant issue is the potential loss of spatial information due to the global nature of the Fourier transform. Furthermore, FFT methods can be computationally demanding, especially for large images, and may perform poorly under non-uniform lighting conditions. These challenges indicate the necessity for further research to optimize FFT-based methods for practical use in low-light image enhancement.

2.3 Retinex Theory

Land and McCann introduced the Retinex theory [20,21] through a series of optical experiments, demonstrating that intrinsic reflectance and incident illumination together determine the radiation reaching the human eye. The mathematical representation is as follows:

$$I = L \circ R \tag{1}$$

Here, the symbol $\circ$ denotes the Hadamard product, where I represents the radiation reaching the human eye, L represents illumination intensity, and R represents reflectance. The reflectance R remains constant for images of the same scene under varying exposure conditions, as it is determined solely by the intrinsic properties of the object's surface. This indicates that color perception primarily depends on reflectance.

Various approaches have utilized the Retinex theory to enhance image quality. For example, numerous studies have employed this theory to improve image quality under different conditions [9,11,33,35]. Other research has refined its application in image processing [3,10,12,16,26,31,39]. Recently, deep learning has become prevalent in the field of low-light image enhancement (LLIE) due to its robust learning capabilities and inference speed. Significant advancements using deep learning techniques have been shown in various studies [8,34,37,38,47]. Moreover, recent research has highlighted different approaches and improvements achieved in LLIE through deep learning [22,28,29,42]. Approximately one-third of these deep learning methods incorporate the Retinex theory to achieve better enhancement effects and provide a physical explanation for the enhancement process [24]. Consequently, leveraging the Retinex theory to guide image enhancement methods in deep learning is crucial for establishing an effective physical model.

3 Proposed Method

The proposed method is structured to address the challenges of enhancing low-light images through an unpaired learning approach, leveraging the inherent properties of the images themselves without reliance on paired high-light images as ground truth. This approach allows for greater flexibility and applicability in practical scenarios where high-light references may not be available. The method is divided into two main components: the Joint Estimation Network and the Multi-Domain Feature Fusion Network, each designed to tackle specific aspects of low-light image enhancement, as shown in Fig. 1.

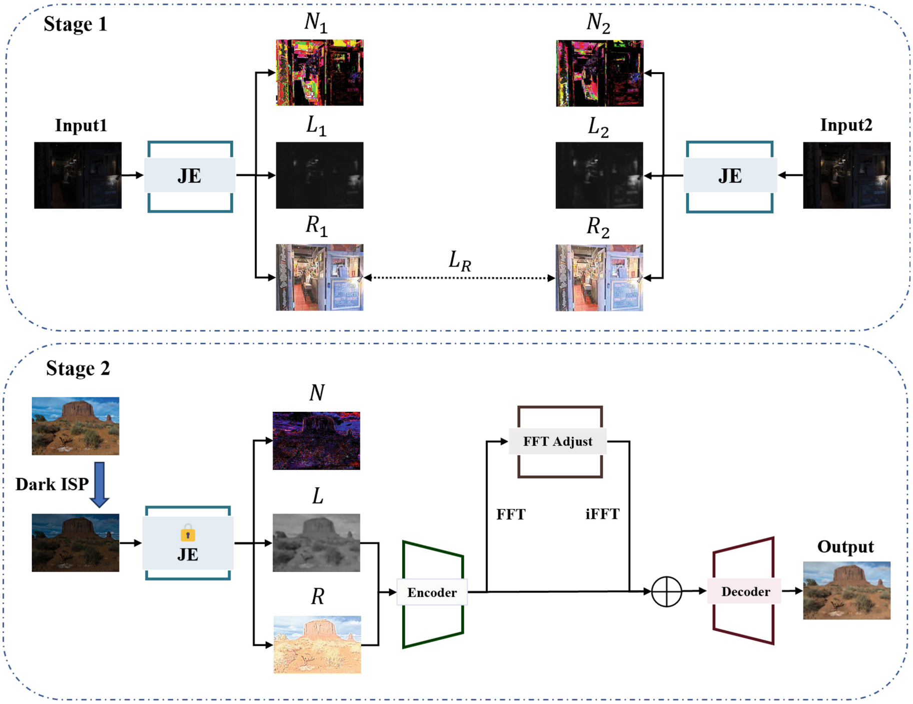


Fig. 1. Overview of the network model structure. The diagram illustrates a two-stage training process. In Stage 1, the Joint Estimation Network(JE) is trained using paired low-light images. In Stage 2, the Joint Estimation Network(JE) is fixed, and the Multi-Domain Feature Fusion Network is trained using another set of normal-light images. R stands for Reflection, L stands for Illuminance, and N stands for Noise.

3.1 Dark ISP

We use the EC-Zero-DCE model [51] to randomly degrade images from normally illuminated inputs. The process involves converting the input images to the LAB color space to isolate the luminance channel.

The core transformation applied to generate low-light images can be expressed as:

$$I_{\text{low}} = I_{\text{orig}} \times \frac{\text{EC-Zero-DCE}(L, E) \times \alpha}{L \times \alpha + \epsilon} \tag{2}$$

where I_{orig} is the original image, EC-Zero-DCE(L, E) is the output of the EC-Zero-DCE model given the luminance channel L and exposure map E, and α is a scaling factor.

The luminance channel L is extracted as follows:

$$L = f_{LAB}(I_{orig}) \tag{3}$$

where f_{LAB} converts the RGB image to LAB color space and extracts the luminance channel.

The exposure map E is defined as:

$$E = \begin{cases} \beta & \text{if } L < \gamma \\ L & \text{otherwise} \end{cases} \tag{4}$$

where γ is a threshold for saturated regions and β is a randomly chosen exposure degree.

To simulate realistic low-light conditions, we add Gaussian noise for sensor noise and JPEG compression artifacts for quality loss. Then, we combine the enhanced low-light luminance channel with the original chrominance channels and convert back to RGB. This ensures the generated low-light images retain the original structure and color, producing high-quality images for training and evaluating low-light image enhancement algorithms.

During the training process, we generate moderately dark and extremely dark images by controlling specific parameter ranges, and we randomly select parameters to degrade the images. These low-light images are then processed through the network separately. We calculate the loss not only between each of these low-light images and the normal light image but also between the two generated low-light images. This ensures that the network effectively learns the features and details of image enhancement under different lighting conditions and enhances images to a uniform level, preventing overexposure. Consequently, this improves the network's performance in enhancing low-light images in real-world scenarios.

3.2 Joint Estimation Network

The Joint Estimation Network is crafted to decompose low-light images into their constituent components of illuminance, reflectance, and noise. This decomposition facilitates a focused enhancement of each attribute, thereby achieving a comprehensive improvement in the overall image quality. The network operates by first estimating the noise within the image and subsequently isolating the illuminance and reflectance components, which are crucial for reconstructing the enhanced image.

The input to this network consists of paired low-light images I_{low}, processed to suppress noise features and enhance underlying details. The architecture effectively models the complex interplay between the different components of the image, using the following decomposition:

$$I_{low} = R \circ L + N \tag{5}$$

where R stands for Reflection, L stands for Illuminance, and N stands for Noise.

In the training phase, we employ a series of loss functions that independently validate the accuracy of each decomposed component. These functions ensure that the network can reconstruct high-quality images by accurately balancing the interdependent relationships between illuminance, reflectance, and noise without the need for high-light ground truth. The specific loss functions used are as follows:

Illuminance Consistency Loss: This loss ensures the estimated illuminance map L closely matches the perceived illuminance of the low-light input, I_{low}. It also incorporates a total variation (TV) loss to smooth the illuminance map [13].

$$\begin{aligned}\mathcal{L}_{\text{illuminance}} = ||L \circ R - I_{low}||_2^2 + ||R - \frac{I_{low}}{L+\epsilon}||_2^2 + ||L - \max(I_{low})||_2^2 \\ + (\|\nabla_h L\|_1 + \|\nabla_w L\|_1)\end{aligned} \tag{6}$$

where ϵ is a small constant to prevent division by zero, $\max(I_{low})$ is the maximum pixel value in I_{low}, and the TV loss term is expanded as $\|\nabla_h L\|_1 + \|\nabla_w L\|_1$, promoting smoothness in the illuminance component.

Reflectance Consistency Loss: This loss ensures that the reflectance component R remains consistent in the same scene, particularly focusing on maintaining texture and color consistency.

$$\mathcal{L}_{\text{reflectance}} = ||R_1 - R_2||_2^2 \tag{7}$$

where R_1 and R_2 are reflectance estimates from different images of the same scene, emphasizing the model's ability to produce stable reflectance maps.

Noise Loss: This loss assesses the effectiveness of the noise reduction by comparing the noise-reduced image I_1 to the original low-light image I_{low}.

$$\mathcal{L}_{\text{noise}} = ||I_{low} - I_1||_2^2 \tag{8}$$

where I_1 represents the image after noise has been processed and reduced by the network.

This unpaired approach underscores the network's adaptability to varied lighting conditions, making it robust for real-world applications.

3.3 Multi-domain Feature Fusion Network

The Multi-Domain Feature Fusion Network adopts a novel approach to enhancing the quality of low-light images by simulating the conditions under which these images might be captured. This network is trained using normal-light images I_{high}, which are processed through a dark Image Signal Processor (ISP) simulation to generate corresponding low-light images $I_{low-sim}$. These simulated images serve as the training input, allowing the network to learn and adapt to a range of low-light environments.

The network architecture is based on a series of encoder and decoder modules that work across multiple scales. The key component, the Bidomain Nonlinear Mapping module [7], extracts spatial features from the input images and then translates these into the frequency domain using Fast Fourier Transform (FFT). Adjustments in the frequency domain focus on enhancing both local details and global consistency, which is crucial for low-light enhancement:

$$X_{output} = X_{spatial} \oplus \mathcal{F}^{-1}(\mathcal{A}(\mathcal{F}(X_{spatial}))) \tag{9}$$

where $X_{spatial}$ is the input image in the spatial domain, $\mathcal{F}$ is the Fourier transform operator, $\mathcal{A}$ is the adjustment function in the frequency domain, $\mathcal{F}^{-1}$ is the inverse Fourier transform, $\oplus$ denotes the feature fusion operation, and X_{output} is the output image.

To ensure the network's effectiveness across various scales, a multi-scale loss function is employed. This function measures the discrepancy between the simulated low-light inputs and the network's outputs, comparing them to normal light images. It incorporates frequency domain losses to ensure a comprehensive enhancement of image quality:

$$L = \lambda_1 \sum_{i=1}^{3} \|\hat{I}_i - I_{high}\|_1 + \lambda_2 \sum_{i=1}^{3} \|\mathrm{FFT}(\hat{I}_i) - \mathrm{FFT}(I_{high})\|_1 \tag{10}$$

where $\hat{I}_i$ represents the network's output image at scale i, I_{high} is the corresponding normal light image used as the ground truth, λ_1 and λ_2 are weighting factors that balance the contribution of spatial domain loss and frequency domain loss, respectively, $\| \cdot \|_1$ denotes the L1 norm, which measures the absolute differences between the predicted and ground truth images, and $\mathrm{FFT}(\cdot)$ represents the Fast Fourier Transform, which transforms the images to the frequency domain.

By using both spatial and frequency domain losses across multiple scales, the network is encouraged to produce outputs that are not only visually similar to the ground truth in terms of pixel values but also consistent in their frequency content, leading to a more comprehensive enhancement of image quality. The input to the network is a simulated low-light image, and the output is compared to the normal light image for evaluation.

4 Experiment

4.1 Experimental Settings

Compared Methods. We compare our methods with model-based method including LIME [16], supervised learning methods including RetinexNet [38], RUAS [28], and ExCNet [45], semi-supervised learning methods including DRBN [43], unpaired supervised learning methods including CLIP-LIT [27], EnlightenGAN [18], and QuadPrior [36], and zero-shot learning methods including RRDNet [52], Zero-DCE [15], Zero-DCE++ [25], NeRCo [41], RetinexDIP [48], SCI [29], and PairLIE [13].

Datasets. We utilized the official test sets of LOL-v1 [38] comprising 15 pairs of low-light and normal-light images and LOL-v2 [43] comprising 100 pairs. Additionally, we followed Retinexformer [5] to split 500 pairs from the MIT-Adobe FiveK dataset [2] for testing. On the LOL and MIT datasets, we reported PSNR, SSIM, and LPIPS [46]. During the first phase of training the Joint Estimation Network, we exclusively use the paired low-light images from the LOL dataset. In the second phase, for the overall network training, we only utilize the normal-light images from the corresponding training datasets. For instance, when training with the FiveK dataset, we only use the normal-light images from the FiveK training set.

Implementations. We use ADAM as the optimizer and employ a learning rate scheduler for learning rate adjustments. The initial learning rate is set to 1×10^{-4} and adjusted every 50 epochs at a decay rate of 0.5. The network is trained for 400 epochs. During training, we crop image patches to (128, 128). The batch size is set to 8. All experiments are conducted on an NVIDIA GeForce RTX 3090 GPU and implemented using PyTorch.

4.2 Quantitative Results

LOL Dataset. Table 1 presents the benchmarking results for low-light enhancement on the LOL-v1 dataset and Table 2 presents the benchmarking results for low-light enhancement on the LOL-v2 dataset. The proposed method achieved the highest scores in PSNR, SSIM, and LPIPS among unsupervised methods, and it was competitive with supervised methods.

MIT-Adobe FiveK Dataset. Table 3 shows the benchmarking results on the MIT-Adobe FiveK dataset. The proposed method achieved competitive performance, particularly in PSNR, SSIM and LPIPS.

4.3 Visual Comparison

Figure 2 and Fig. 3 show a visual comparison of the proposed method with other methods on the MIT-Adobe FiveK and LOL-v1 dataset. The proposed method generates images with better illumination, color consistency, and naturalness.

Table 1. Benchmarking results for low-light enhancement on the LOL-v1 dataset. The highest scores are highlighted in red, and the second highest scores are marked in blue.

	Input	LIME [16]	RetinexNet [38]	RUAS [28]	DRBN [43]
PSNR↑	7.77	16.76	16.77	16.40	15.13
SSIM↑	0.181	0.560	0.462	0.537	0.472
LPIPS↓	0.560	0.350	0.474	0.350	0.316
	EnlightenGAN [18]	RRDNet [52]	RetinexDIP [48]	ZeroDCE [15]	ZeroDCE++ [25]
PSNR↑	17.48	11.38	11.65	14.86	15.34
SSIM↑	0.677	0.470	0.501	0.589	0.603
LPIPS↓	0.322	0.361	0.317	0.335	0.316
	SCI [29]	PairLIE [13]	QuadPrior [36]	Ours	
PSNR↑	14.78	18.46	18.34	20.91	
SSIM↑	0.553	0.749	0.827	0.773	
LPIPS↓	0.332	0.290	0.209	0.261	

Table 2. Benchmarking results for low-light enhancement on the LOL-v2 dataset.

	Input	LIME [16]	RetinexNet [38]	RUAS [28]	DRBN [43]
PSNR↑	9.72	15.24	15.47	15.33	19.60
SSIM↑	0.190	0.470	0.560	0.520	0.764
LPIPS↓	0.333	0.360	0.421	0.322	0.246
	EnlightenGAN [18]	RRDNet [52]	RetinexDIP [48]	ZeroDCE [15]	ZeroDCE++ [25]
PSNR↑	18.23	14.85	14.51	18.06	18.49
SSIM↑	0.610	0.560	0.546	0.605	0.617
LPIPS↓	0.309	0.265	0.274	0.298	0.290
	SCI [29]	PairLIE [13]	QuadPrior [36]	Ours	
PSNR↑	17.30	19.89	20.31	20.44	
SSIM↑	0.565	0.778	0.808	0.780	
LPIPS↓	0.286	0.282	0.202	0.264	

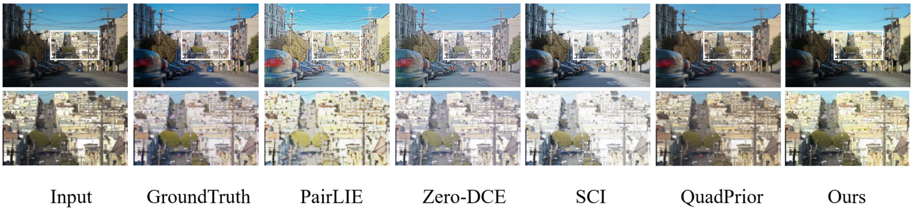
4.4 Ablation Studies

To thoroughly evaluate the contributions of each component in our proposed model, we conducted ablation studies on the LOL-v1 dataset. Specifically, we removed certain modules from the full model to understand their impact on the overall performance. Here, w/o Multi-Domain Feature Fusion refers to connecting a decoder directly after the Joint Estimation Network. The quantitative results are presented in Table 4.

From these ablation studies, it is evident that both the Joint Estimation Network and the Multi-Domain Feature Fusion module play crucial roles in our model. The Multi-Domain Feature Fusion significantly contributes to the perceptual quality and accurate reconstruction of the images, as evidenced by the LPIPS and PSNR metrics. On the other hand, the Joint Estimation Network is

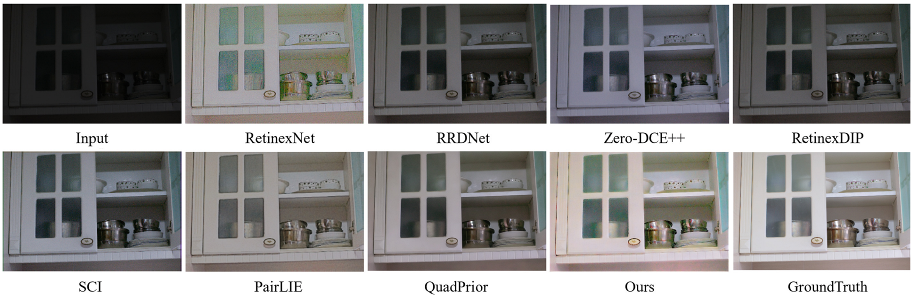
Table 3. Benchmarking results for low-light enhancement on the MIT-Adobe FiveK dataset.

	ExCNet [45]	EnlightenGAN [18]	PairLIE [13]	NeRCo [41]
PSNR↑	14.21	13.28	10.55	17.33
SSIM↑	0.719	0.738	0.642	0.767
LPIPS↓	0.197	0.203	0.273	0.208
	CLIP-LIT [27]	ZeroDCE [15]	ZeroDCE++ [25]	RUAS [28]
PSNR↑	17.00	13.53	12.33	9.53
SSIM↑	0.781	0.725	0.408	0.610
LPIPS↓	0.159	0.201	0.280	0.301
	SCI [29]	QuadPrior [36]	Ours	
PSNR↑	16.29	18.51	20.90	
SSIM↑	0.795	0.785	0.833	
LPIPS↓	0.143	0.163	0.163	



Input GroundTruth PairLIE Zero-DCE SCI QuadPrior Ours

Fig. 2. Example low-light enhancement results on the MIT-Adobe FiveK.



Input RetinexNet RRDNet Zero-DCE++ RetinexDIP

SCI PairLIE QuadPrior Ours GroundTruth

Fig. 3. Example low-light enhancement results on the LOL-v1 dataset.

Table 4. Ablation study results on the LOL-v1 dataset.

Model Variant	PSNR ↑	SSIM ↑	LPIPS ↓
Full Model	20.91	0.773	0.261
w/o Joint Estimation Network	19.19	0.718	0.359
w/o Multi-Domain Feature Fusion	17.59	0.793	0.385

essential for capturing fine structural details, thereby improving the SSIM and ensuring better overall image quality. These findings validate the design choices made in our proposed model, demonstrating their effectiveness in enhancing low-light image enhancement tasks.

5 Conclusion

In this paper, we proposed a novel unpaired method for low-light image enhancement that leverages Retinex theory and Fast Fourier Transform (FFT)-based processing. The method consists of a Joint Estimation Network and a Multi-Domain Feature Fusion Network, which decompose low-light images into illuminance, reflectance, and noise components, and integrate spatial and frequency domain information to enhance image details and maintain global consistency. Extensive experiments on public datasets demonstrate that our approach significantly improves the visual quality of low-light images, outperforming existing unsupervised and zero-shot learning methods in both qualitative and quantitative metrics. The key advantages of our approach include the ability to enhance images without paired training data, making it highly applicable in real-world scenarios. Future work could explore integrating additional domain adaptation techniques and extending the framework to handle video sequences, providing further benefits for applications in surveillance, autonomous driving, and low-light photography.

Acknowledgment. This work was supported by the innovation fund (No. 2021YFB3601400). I would like to express my sincere gratitude to my advisors, Prof. Fanjiang Xu and Prof. Xiongxin Tang, for their invaluable guidance and support throughout this research. I also extend my thanks to Prof. Hanxiang Yang, and Prof. Chunle Guo for their insightful feedback and assistance in the development of this work. Furthermore, I am deeply grateful to the development team of the SEE series software for providing the essential simulation support that greatly contributed to the completion of this research.

References

1. Afifi, M., Derpanis, K.G., Ommer, B., Brown, M.S.: Learning multi-scale photo exposure correction. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 9157–9167 (2021)
2. Bychkovsky, V., Paris, S., Chan, E., Durand, F.: Learning photographic global tonal adjustment with a database of input/output image pairs. In: CVPR (2011)
3. Cai, B., et al.: A joint intrinsic-extrinsic prior model for retinex. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 4000–4009 (2017)
4. Cai, J., Gu, S., Zhang, L.: Learning a deep single image contrast enhancer from multi-exposure images. IEEE Trans. Image Process. **27**(4), 2049–2062 (2018). https://doi.org/10.1109/TIP.2017.2786696
5. Cai, Y., Bian, H., In, J., Wang, H., Timofte, R., Zhang, Y.: Retinexformer: one-stage retinex-based transformer for low-light image enhancement. In: ICCV (2023)

6. Chen, C., Chen, Q., Xu, J., Koltun, V.: Learning to see in the dark. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 3291–3300 (2018)
7. Cong, X., Gui, J., Zhang, J., Hou, J., Shen, H.: A semi-supervised nighttime dehazing baseline with spatial-frequency aware and realistic brightness constraint (2024). https://arxiv.org/abs/2403.18548
8. Fan, M., et al.: 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)
9. Fu, X., et al.: A novel retinex based approach for image enhancement with illumination adjustment. In: 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 1190–1194. IEEE (2014)
10. Fu, X., et al.: A fusion-based enhancing method for weakly illuminated images. Sig. Process. **129**, 82–96 (2016)
11. Fu, X., et al.: A weighted variational model for simultaneous reflectance and illumination estimation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2782–2790 (2016)
12. Fu, X., et al.: Retinex-based perceptual contrast enhancement in images using luminance adaptation. IEEE Access **6**, 61277–61286 (2018)
13. Fu, Z., Yang, Y., Tu, X., Huang, Y., Ding, X., Ma, K.: Learning a simple low-light image enhancer from paired low-light instances. In: CVPR (2023)
14. Fuoli, D., Gool, L.V., Timofte, R.: Fourier space losses for efficient perceptual image super-resolution. IEEE Conference on Computer Vision and Pattern Recognition, pp. 2360–2369 (2021)
15. Guo, C.G., et al.: Zero-reference deep curve estimation for low-light image enhancement. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1780–1789 (2020)
16. Guo, X., Li, Y., Ling, H.: Lime: low-light image enhancement via illumination map estimation. IEEE Trans. Image Process. **26**(2), 982–993 (2016)
17. Huang, J., et al.: Deep Fourier-based exposure correction network with spatial-frequency interaction. In: Avidan, S., Brostow, G., Cissé, M., Farinella, G.M., Hassner, T. (eds.) ECCV 2022. LNCS, vol. 13679, pp. 163–180. Springer, Cham (2022). https://doi.org/10.1007/978-3-031-19800-7_10
18. Jiang, Y., et al.: EnlightenGAN: deep light enhancement without paired supervision. IEEE TIP **30**, 2340–2349 (2021)
19. Jobson, D.J., Rahman, Z., Woodell, G.A.: A multiscale retinex for bridging the gap between color images and the human observation of scenes. IEEE TIP **6**(7), 965–976 (1997)
20. Land, E.H.: The retinex theory of color vision. Sci. Am. **237**(6), 108–129 (1977)
21. Land, E.H., McCann, J.J.: Lightness and retinex theory. JOSA **61**(1), 1–11 (1971)
22. Li, C., et al.: Lightennet: a convolutional neural network for weakly illuminated image enhancement. Pattern Recogn. Lett. **104**, 15–22 (2018)
23. Li, C., et al.: Embedding Fourier for ultra-high-definition low-light image enhancement. In: ICLR (2023)
24. Li, C., et al.: Low-light image and video enhancement using deep learning: a survey. IEEE Trans. Pattern Anal. Mach. Intell. **44**(12), 9396–9416 (2021)
25. Li, C., Guo, C., Loy, C.C.: Learning to enhance low-light image via zero-reference deep curve estimation. IEEE TPAMI **44**(8), 4225–4238 (2021)
26. Li, M., et al.: Structure-revealing low-light image enhancement via robust retinex model. IEEE Trans. Image Process. **27**(6), 2828–2841 (2018)
27. Liang, Z., Li, C., Zhou, S., Feng, R., Loy, C.C.: Iterative prompt learning for unsupervised backlit image enhancement. In: ICCV (2023)

28. Liu, R., Ma, L., Zhang, J., Fan, X., Luo, Z.: Retinex-inspired unrolling with cooperative prior architecture search for low-light image enhancement. In: CVPR (2021)
29. Ma, L., Ma, T., Liu, R., Fan, X., Luo, Z.: Toward fast, flexible, and robust low-light image enhancement. In: CVPR (2022)
30. Pizer, S.M., Johnston, R.E., Ericksen, J.P., Yankaskas, B.C., Muller, K.E.: Contrast-limited adaptive histogram equalization: speed and effectiveness. In: VBC (1990)
31. Ren, W., et al.: Low-light image enhancement via a deep hybrid network. IEEE Trans. Image Process. **28**, 4364–4375 (2019)
32. Wang, C., Wu, H., Jin, Z.: Fourllie: boosting low-light image enhancement by Fourier frequency information. arXiv preprint arXiv:2308.03033 (2023)
33. Wang, L., et al.: Variational Bayesian method for retinex. IEEE Trans. Image Process. **23**(8), 3381–3396 (2014)
34. Wang, R., et al.: Underexposed photo enhancement using deep illumination estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 6849–6857 (2019)
35. Wang, S., et al.: Naturalness preserved enhancement algorithm for non-uniform illumination images. IEEE Trans. Image Process. **22**(9), 3538–3548 (2013)
36. Wang, W., Yang, H., Fu, J., Liu, J.: Zero-reference low-light enhancement via physical quadruple priors (2024)
37. Wang, Y., et al.: 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)
38. Wei, C., Wang, W., Yang, W., Liu, J.: Deep retinex decomposition for low-light enhancement. In: BMVC (2018)
39. Xu, J., et al.: Star: a structure and texture aware retinex model. IEEE Trans. Image Process. **29**, 5022–5037 (2020)
40. Xu, Q., Zhang, R., Zhang, Y., Wang, Y., Tian, Q.: A Fourier-based framework for domain generalization. IEEE Conference on Computer Vision and Pattern Recognition, pp. 14383–14392 (2021)
41. Yang, S., Ding, M., Wu, Y., Li, Z., Zhang, J.: Implicit neural representation for cooperative low-light image enhancement. In: ICCV (2023)
42. Yang, W., et al.: Sparse gradient regularized deep retinex network for robust low-light image enhancement. IEEE Trans. Image Process. **30**, 2072–2086 (2021)
43. 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: CVPR (2020)
44. Yu, H., Zheng, N., Zhou, M., Huang, J., Xiao, Z., Zhao, F.: Frequency and spatial dual guidance for image dehazing. In: Avidan, S., Brostow, G., Cissé, M., Farinella, G.M., Hassner, T. (eds.) ECCV 2022. LNCS, vol. 13679, pp. 181–198. Springer, Cham (2022). https://doi.org/10.1007/978-3-031-19800-7_11
45. Zhang, L., Zhang, L., Liu, X., Shen, Y., Zhang, S., Zhao, S.: Zero-shot restoration of back-lit images using deep internal learning. In: ACM MM (2019)
46. 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)
47. Zhang, Y., et al.: Kindling the darkness: a practical low-light image enhancer. In: Proceedings of the 27th ACM International Conference on Multimedia, pp. 1632–1640 (2019)

48. Zhao, Z., Xiong, B., Wang, L., Ou, Q., Yu, L., Kuang, F.: Retinexdip: a unified deep framework for low-light image enhancement. IEEE Trans. Circuits Syst. Video Technol. **32**(3), 1076–1088 (2022). https://doi.org/10.1109/TCSVT.2021.3073371
49. Zhou, M., et al.: Adaptively learning low-high frequency information integration for pan-sharpening. In: ACM International Conference on Multimedia, pp. 3375–3384 (2022)
50. Zhou, M., et al.: Deep Fourier up-sampling. In: Advances in Neural Information Processing Systems, vol. 35, pp. 22995–23008 (2022)
51. Zhou, S., Li, C., Change Loy, C.: LEDNet: joint low-light enhancement and deblurring in the dark. In: Avidan, S., Brostow, G., Cissé, M., Farinella, G.M., Hassner, T. (eds.) ECCV 2022. LNCS, vol. 13666, pp. 573–589. Springer, Cham (2022). https://doi.org/10.1007/978-3-031-20068-7_33
52. Zhu, A., Zhang, L., Shen, Y., Ma, Y., Zhao, S., Zhou, Y.: Zero-shot restoration of underexposed images via robust retinex decomposition. In: ICME (2020)

PSTNet: A Progressive Sparse Transformer Network for Image Deraining

Raj Ahamed Shaik, Navjot Singh(✉), Antony Verriboina,
and Debesh Kumar Shandilya

Computer Vision and Biometrics Lab, Department of Information Technology,
Indian Institute of Information Technology, Allahabad, Uttar Pradesh, India
navjot@iiita.ac.in

Abstract. Image deraining aims to transform a rainy input image into an image of high quality. Transformer-based techniques have demonstrated remarkable efficacy in image deraining because of their capacity to represent non-local information, a crucial element for high-quality image reconstruction. Our findings indicate that most transformers used for feature aggregation nowadays leverage all token similarities between query-key pairs. If the query tokens differ from the key tokens, the quality of the recovered image is compromised by the self-attention values derived from these tokens if these are considered during feature aggregation. For this purpose, we propose the Progressive Sparse Transformer Network (PSTNet), which progressively learns to restore degraded inputs by retaining the helpful self-attention information during feature aggregation and discarding the remaining values which obstruct the restoration. Transformer blocks help to capture interactions between distant pixels. Thorough experimental outcomes on widely used benchmarks show that the suggested approach performs better than most existing techniques.

1 Introduction

Single image deraining is a prevalent ill-posed vision challenge that has surfaced in the recent decade, which makes an effort to reconstruct a clear output from the rain-degraded input. The unknown rain streak pattern constitutes a hard situation that requires solid image priors for effective restoration. To address this issue, early techniques [25,29,67] usually impose different priors depending on the statistical features of rain streaks. The deraining performance of these handcrafted priors is limited since they cannot withstand complex and varied rainy settings.

Convolutional neural networks (CNN) obtain generalizable priors from vast amounts of data, making them a better choice than traditional restoration approaches. The fundamental process of CNNs is 'convolution', which provides local connection and translation equivariance. These characteristics increase CNNs' effectiveness and generality, but they also present two significant challenges: (a) Because of its small receptive field, the convolution operator cannot

A. Antonacopoulos et al. (Eds.): ICPR 2024, LNCS 15332, pp. 106–120, 2025.
https://doi.org/10.1007/978-3-031-78125-4_8

describe long-range pixel interactions, and (b) The static weights of convolution filters restrict their capacity to adjust to input data.

Transformers [5,30,39,54] have been used for image deraining to get over these restrictions, and they've done an admirable job of it since they can more accurately represent the non-local relations required to rebuild images with high quality. However, when clear images are restored, these techniques fall short of accurately simulating the localized characteristics of images. Transformers' self-attention fails to model the local invariant features that CNNs excel at, which is one of the key causes. In local regions, rain streaks are often mistaken for background. To overcome these constraints, recent works [10,23,61] combine CNN operations with transformers.

Conventional transformers [46] consider all attention values based on query-key pairs to aggregate features. Sometimes, as the tokens from the key tokens may not always be relevant to the query tokens, applying the self-attention values computed from these tokens may hinder the process when reconstructing the output. The reason is that smaller similarity weights are often amplified by the dense computation pattern of self-attention, allowing implicit noise into the feature interaction and aggregation process. Consequently, when modelling global feature dependencies, redundant or unnecessary representations are frequently included [48,73]. These realizations motivate us to determine and use the best self-attention values to maximize feature utilization for enhanced image restoration.

To overcome the difficulties, we have developed PSTNet, an efficient progressive sparse Transformer network to restore the image. The core of this framework is the selective multi-head attention (SMHA) and a simple Gate Feed-forward network (SGFN). The SMHA mechanism replaces the traditional self-attention by retaining only the k most crucial similarity scores between queries and keys, thus enhancing feature aggregation, and the remaining scores are discarded. The gating mechanism in SGFN regulates the flow of complementary features, enabling subsequent network layers to concentrate on more refined image attributes. Three main benefits come with our suggested method: (1) better robustness because of less sensitivity to irrelevant feature interference; (2) better localization and global feature utilization; and (3) better deraining performance by utilizing both data and content sparsity.

An overview of the primary contributions is provided below.:

- We introduce a sparse Transformer architecture to achieve better deraining results with improved detail and texture recovery.
- We propose using SMHA as part of a sparse transformer in the encoder-decoder architecture. This mechanism is intended to gather the most crucial data from the collected feature maps.
- A novel, simple gate feed-forward network (SGFN) that regulates feature transformation by filtering out less informative features in the network has been developed.
- Comprehensive tests on multiple benchmarks demonstrate that our model performs over state-of-the-art (SOTA) techniques.

2 Related Work

2.1 Single Image Deraining

Conventional techniques [19,25,29,35,67] for image deraining frequently create an image-prior to impose extra constraints, but these manually created priors are dependent on empirical observations and find it difficult to capture the intrinsic characteristics of distinct images. Many frameworks based on CNN have been developed to tackle it [57], and their performance greatly surpassed that of older approaches. By taking into account attributes like rain direction [32], density [68], and veiling effect [20], as well as by optimizing network architectures through the use of transfer mechanisms [21,53,59,60] or recursive computation [24,28,41], certain studies have improved the depiction of rain. Despite their achievements, the constraints of convolution make it difficult for these algorithms to capture long-range dependencies. Because of its computational efficiency and hierarchical multi-scale representation, encoder-decoder-based U-Net architectures [1,11,27,51,63,66,71] are very popular. Furthermore, skip connection-based methods that concentrate on residual signal learning [18,31,65,72] have shown effectiveness. Selectively attending to relevant details [28,65,66] has also benefited from integrating spatial and channel attention modules. We can effectively simulate non-local information by employing a transformer as the network's backbone.

2.2 Vision Transformers

Transformers were initially created for challenges involving the processing of sequences in natural language, [46] and have since been adapted for various vision tasks, including image recognition [13,45,62], detection [3,33,75] and segmentation [49,55,74]. Vision Transformers break down an image into a series of patches and discover how they relate to each other [13,45], providing a strong capability to understand long-range relationships and adjust to input data [26]. These characteristics have led to their application in image deraining [51]. Jiang et al. [23] combined a background restoration network with self-attention in a Transformer to create a dynamic deraining network. Recently, Xiao et al. [54] introduced the image deraining Transformer (IDT), which uses a dual Transformer approach combining spatial and window-based attention to get outstanding outcomes. However, most of the current approaches rely on all the self-attention scores, which can include redundant or irrelevant features with smaller weights, leading to potential noise in the output features. To address it, we propose using sparse attention in Transformers to focus on the most relevant information and reduce noise.

2.3 Sparse Attention

Inspired by the neural activity in biological brains, the concept of sparsity in hidden representations within deep neural networks offers significant advantages

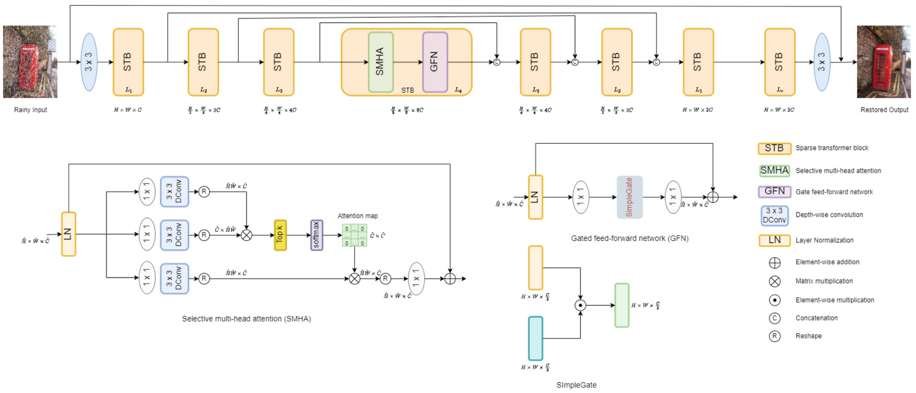

Rainy Input
3 x 3
STB
L_1
$H \times W \times C$
STB
L_2
$\frac{H}{2} \times \frac{W}{2} \times 2C$
STB
L_3
$\frac{H}{4} \times \frac{W}{4} \times 4C$
SMHA
GFN
STB
L_4
$\frac{H}{8} \times \frac{W}{8} \times 8C$
©
STB
L_3
$\frac{H}{4} \times \frac{W}{4} \times 4C$
©
STB
L_2
$\frac{H}{2} \times \frac{W}{2} \times 2C$
©
STB
L_1
$H \times W \times 2C$
STB
L_r
$H \times W \times 2C$
3 x 3
Restored Output
$\hat{H} \times \hat{W} \times \hat{C}$
LN
1 x 1
3 x 3 DConv
R
$\hat{H}\hat{W} \times \hat{C}$
$\hat{C} \times \hat{H}\hat{W}$
Top k
softmax
Attention map
0 ... 0
0 ... 0
$\hat{C} \times \hat{C}$
$\hat{H}\hat{W} \times \hat{C}$
$\hat{H} \times \hat{W} \times \hat{C}$
Selective multi-head attention (SMHA)
$\hat{H} \times \hat{W} \times \hat{C}$
LN
1 x 1
SimpleGate
1 x 1
$\hat{H} \times \hat{W} \times \hat{C}$
Gated feed-forward network (GFN)
$H \times W \times \frac{C}{2}$
$H \times W \times \frac{C}{2}$
$H \times W \times \frac{C}{2}$
SImpleGate
STB Sparse transformer block
SMHA Selective multi-head attention
GFN Gate feed-forward network
3 x 3 DConv Depth-wise convolution
LN Layer Normalization
Element-wise addition
Matrix multiplication
Element-wise multiplication
Concatenation
Reshape

Fig. 1. Architecture of Progressive sparse transformer network (PSTNet) for image deraining. The primary modules of STB are selective multi-head attention (SMHA), which masks out unwanted information, and a simple gate feed-forward network (SGFN) with a simple gate for useful information to propagate further.

for both problems related to NLP and vision [48,73]. Sparse representation addresses low-level vision issues like super-resolution [37] and image draining [50]. Sparse attention mechanisms can be classified as content-driven sparse attention and fixed (data-driven) sparse attention [9,12,43]. Data-driven sparse attention often involves introducing local attention operations into a CNN backbone, focusing primarily on local window sizes. Recent studies [17,47] have explored enforcing sparsity in Transformer architectures, such as Zhang et al.'s [70] attention retractable Transformer, which allows interaction among features from sparse areas. Unlike these approaches, we implement a straightforward yet effective approximation for self-attention based on the most crucial attention values.

2.4 Selective Attention

For NLP tasks, Zhao et al. were the first to provide an explicit selection strategy based on the most crucial attention values. Vision transformer has been improved with the introduction of k-NN attention, building on its success. We have created an effective SMHA, contrasting to the selective attention used in the spatial dimension.

3 Proposed Model

First, as illustrated in Fig. 1, we provide the overall flow of our PSTNet architecture. Next, we outline the essential elements of the suggested sparse transformer: (a) Selective multi-head attention (SMHA) and (b) SimpleGate feed-forward network (SGFN). Lastly, we provide a progressive training strategy.

3.1 Overall Pipeline

The flow of our proposed PSTNet, as seen in Fig. 2, a modified U-net encoder-decoder design is used. Given a rainy image $I_{rain} \in \mathbb{R}^{H \times W \times 3}$; where C represents channel count and H×W is the spatial resolution. First, to acquire low-level features $F \in \mathbb{R}^{H \times W \times C}$, PSTNet applies convolution. These shallow features F_0 are converted into deep features $D \in \mathbb{R}^{H \times W \times 2C}$ after passing through a symmetric encoder-decoder network. At every stage, the encoder-decoder targets a different spatial resolution and channel dimension to extract multi-scale representations from rainy images. Encoder is used to increase channel capacity by reducing the spatial size hierarchically, beginning with the high-resolution input. Next, the decoder gradually restores the high-resolution representations from low-resolution input (latent features) $L \in \mathbb{R}^{\frac{H}{8} \times \frac{W}{8} \times 8C}$. We employ pixel-unshuffle and pixel-shuffle methods [44] for downsampling and upsampling features, respectively.

Skip connections [42] is used to concatenate encoder and decoder characteristics to facilitate recovery. After concatenation, a 1×1 convolution reduces channels by a scale of 0.5 at all levels except level one. At the top level, sparse transformer blocks combine low-level details from the encoder with high-level details from the decoder, which helps preserve subtle structural and textural characteristics in the recovered images. In each sparse transformer, we introduce SMHA instead of the standard self-attention to achieve feature sparsity, enhancing the feature aggregation process. Additionally, we incorporate a SGFN into the block, aiding in image restoration. This hybrid approach allows PSTNet to leverage both intrinsic properties and the adaptive content of rainy images, effectively separating unwanted rain streaks from the latent clear background. Experiments confirm that these design choices lead to improved image quality. Finally, the refined features are passed through a convolution layer to obtain a residual image $I_{residual} \in \mathbb{R}^{H \times W \times 3}$, it is added to the degraded input to obtain the clean output $I_{clean} = I_{rain} + I_{residual}$. The network is trained to minimize the error function:

$$E = \|I_{clean} - I_{rain}\|_1 \tag{1}$$

where $\|.\|_1$ is L_1-norm. Now, we present the components of the sparse Transformer.

3.2 Sparse Transformer

Standard transformers [13,46,64] compute self-attention globally across all tokens, which can lead to noisy interactions between irrelevant features, making them less effective for image deraining. To resolve the problem, we introduce a sparse transformer for feature extraction, leveraging the benefits of sparsity found in neural networks. For the input features from the (t-1)-th block F_{t-1}, the encoding process of the sparse transformer can be formally described as follows:

$$F'_t = F_{t-1} + \text{SMHA}(\text{LN}(F_{t-1})) \tag{2}$$

$$F_t = F'_t + \text{SGFN}(\text{LN}(F'_t)) \tag{3}$$

where LN stands for layer normalization; F'_t and F_t are the outputs from the selective multi-head self-attention (SMHA) and the simple Gate feed-forward network (SGFN) respectively, as described below.

Selective Multi-head Self-attention (SMHA). Let us revisit the typical self-attention mechanism used in Transformers prevalent in many existing models. In typical attention, given matrices Q (query), K (key), and V (value) with dimensions $\mathbb{R}^{L\times d}$, the output is:

$$\mathrm{A}(Q, K, V) = \sigma\left(\frac{QK^T}{\lambda}\right)V, \tag{4}$$

where σ represents the softmax function. Here, λ is an optional scaling factor defined as $\lambda = \sqrt{d}$. Generally, multi-head attention computes k separate Q, K, and V matrices for each head, which gives $d = C/k$ dimensional results per head. The final result is then obtained across all heads by concatenating and linearly projecting the outputs. The main computational challenge in Transformers is the self-attention layer. In conventional self-attention mechanisms [13,46], the time and memory complexity of taking the dot-product between keys and queries grows quadratically as input spatial resolution increases, specifically $O(W^2\, H^2)$ for $W \times H$ images. Therefore, applying self-attention to many image restoration tasks involving high-resolution images becomes impractical due to these computational demands. So instead of computing self-attention (SA) over spatial dimensions [64], we apply over channels. This involves calculating cross-covariance between channels to produce an attention map with a linear time complexity that implicitly captures the global context. Our strategy prevents irrelevant information from being included throughout the feature interaction phase by replacing previous methods with SMHA.

To capture channel-wise spatial context, we first use 1×1 convolutions to integrate cross-channel context by each pixel, followed by 3×3 depth-wise convolutions. Next, self-attention can be determined across channels. After, the similarity between pairs of pixels is calculated using reshaped queries and keys. Next, we eliminate elements with lower attention weights in the transposed attention map M with size $\mathbb{R}^{\hat{C}\times\hat{C}}$. We choose the k most contributive scores from M using an adaptive method instead of a dropout strategy that arbitrarily discards results. This approach aims to keep the most important elements and eliminate the less beneficial ones [8,9]. Here, k is a parameter that dynamically regulates the sparsity level that can be adjusted. Specifically, it is determined by averaging weighted fractions, such as $\frac{1}{2}$ or $\frac{2}{3}$. Elements in M that do not rank among the k highest scores are not considered when computing probabilities. The sparse attention can be derived as:

$$\mathrm{SparseAttention}(Q, K, V) = \sigma\left(\mathrm{Hk}\left(\frac{QK^T}{\lambda}\right)\right)V, \tag{5}$$

here Hk(·) can be learnt and selects highest k values:

$$[\text{Hk}(A)]_{ij} = \begin{cases} A_{ij} & \text{if } A_{ij} \in \text{highest k values(row } j) \\ 0 & \text{otherwise} \end{cases} \tag{6}$$

Finally, we multiply the matrix to combine the softmax output and the value matrix. Using the multi-head technique, we concatenate the outputs from each attention head and apply a linear projection to get the desired outcome.

SimpleGate Feed-Forward Network. A typical feed-forward network (FN) [13,46] handles each pixel location evenly and separately to modify features. It uses two 1×1 convolutions: the first decreases the channels to the original input size, and the second increases the number of feature maps (usually by a factor of $\gamma = 4$). In the hidden layer, a non-linear activation function is used. Recently, efforts have been made to include a gating mechanism where two parallel channels of linearly transformed layers and non-linearity are induced in one of them (usually GELU) [64]. GELU may be seen as a variant of a Gated Linear Unit (GLU). GLU is formulated as follows:

$$\text{GLU}(\mathbf{Z}, \sigma, f, g) = f(\mathbf{Z}) \odot \sigma(g(\mathbf{Z})) \tag{7}$$

It is possible to think of GLU as an extension of activation functions, with the potential to replace nonlinear activation functions [6]. It is observed that nonlinearity exists in the GLU itself without σ: $GLU(Z) = f(Z) \odot g(Z)$ contains nonlinearity even in the absence of σ, which is termed as SimpleGate, and it is formulated as:

$$\text{simpleGate}(\mathbf{Z}, f, g) = f(\mathbf{Z}) \odot g(\mathbf{Z}) \tag{8}$$

where M and N are identically sized feature maps. It simply divides the input into two equal parts across channel dimensions and multiplies them. For an input tensor $\mathbf{Z} \in \mathbb{R}^{\hat{H} \times \hat{W} \times \hat{C}}$, SGFN is represented as:

$$\hat{\mathbf{Z}} = \text{SimpleGate}(\text{LN}(\mathbf{Z})) + \mathbf{Z} \tag{9}$$

where LN is layer normalization [2]. By controlling the flow of information across the many hierarchical stages of our pipeline, the SGFN allows each step to focus on specific aspects that enhance the work of the other stages. This characteristic sets SGFN apart from SMHA, which is mostly focused on integrating contextual data with features enhancing them.

3.3 Progressing Learning

Typical training for CNN-based restoration models uses fixed-size patches in an image. Nevertheless, suppose a transformer model is trained on small patches;

Table 1. Dataset details showing samples count for image deraining.

Datasets	Rain14000 [15]	Rain12 [29]	Rain1800 [56]	Rain100L [56]	Rain100H [56]	Rain1200 [68]	Rain800 [69]
Train	11200	12	1800	0	0	0	700
Test	2800	0	0	100	100	1200	100
Testset name	Test2800	-	-	Rain100L	Rain100H	Test1200	Test100

it may not be able to acquire global image statistics well enough, which could lead to less-than-ideal performance when tested on full-resolution photos. We use progressive learning to address this, gradually increasing the patch size in later epochs after beginning with smaller image patches in the earlier epochs. This method improves performance while testing with different-resolution images, which is typical for image deraining. Like curricular learning, progressive learning ensures fine image structure and texture preservation by starting the network with easier tasks and working on more difficult ones. To maintain constant optimization time, the batch size is decreased as the patch size grows.

4 Experiments and Analysis

Our model is trained using 13,712 rainy-clean image pairs from several datasets [15,29,56,56,56,68,69], as shown in Table 1. We assess the proposed PSTNet for image deraining on the datasets listed.

Evaluation Metrics. Evaluation measures that are frequently used in deraining benchmarks include PSNR [22] and SSIM [52]. Similar to previous deraining methods [16,24], we calculate SSIM and PSNR measures in YCbCr colour space.

Implementation Details. Our PSTNet architecture is a four-level deep encoder-decoder structure. The number of sparse transformers for levels 1 through 4 is [4, 6, 6, 8], while the attention heads in SMHA are [1, 2, 4, 8]. The original channel count (C) is 32, and 2 is the expansion factor. The refinement stage comprises four blocks. The sparsity parameters for STB in SMHA are set to $\left[\frac{1}{2}, \frac{4}{5}\right]$. Models are trained with the AdamW optimizer ($\beta_1 = 0.9$, $\beta_2 = 0.999$, weight decay $= 1{\times}10^{-4}$) and L1 loss for 300K iterations. The initial learning rate is $3{\times}10^{-4}$, which gradually decreases to $1{\times}10^{-6}$ using cosine annealing [34]. Data augmentation also includes random flips both horizontally and vertically. Data augmentation includes random vertical and horizontal flips.

Image Deraining Results. Table 2 demonstrates that our PSTNet consistently outperforms existing methods across five datasets, delivering performance improvements. When compared to the latest method, Restormer [64], PSTNet shows an average improvement of 0.47 dB across the datasets. Figure 2 presents a visual example where our PSTNet successfully generates a rain-free image while maintaining the structural details effectively. Table 3 shows that our model is more efficient and has fewer parameters and MACs.

Table 2. Comparison of results across five datasets. The best and second-best results are highlighted in bold and underlined, respectively.

Method	Test100		Rain100L		Rain100H		Test1200		Test2800		Average	
	PSNR↑	SSIM↑	PSNR↑	SSIM↑	PSNR↑	SSIM↑	PSNR↑	SSIM↑	PSNR↑	SSIM↑	PSNR↑	SSIM↑
DerainNet [14]	22.77	0.810	27.03	0.884	14.92	0.592	23.38	0.835	24.31	0.861	22.48	0.796
SEMI [53]	22.35	0.788	25.03	0.842	16.56	0.486	26.05	0.822	24.43	0.782	22.88	0.744
DIDMDN [68]	22.56	0.818	25.23	0.741	17.35	0.524	29.65	0.901	28.13	0.867	24.58	0.770
UMRL [58]	24.41	0.829	29.18	0.923	26.01	0.832	30.55	0.910	29.97	0.905	28.02	0.880
RESCAN [28]	25.00	0.835	29.80	0.881	26.36	0.786	30.51	0.882	31.29	0.904	28.59	0.857
GCANet [4]	24.93	0.846	30.63	0.892	26.45	0.783	30.49	0.881	31.42	0.882	29.58	0.856
PreNet [41]	24.81	0.851	32.44	0.950	26.77	0.858	31.36	0.911	31.75	0.916	29.42	0.897
MSPFN [24]	27.50	0.876	32.40	0.933	28.66	0.860	32.39	0.916	32.82	0.930	30.75	0.903
MRPNet [66]	30.27	0.897	36.40	0.965	30.41	0.890	32.91	0.916	33.36	0.926	32.67	0.919
SPAIR [38]	30.35	**0.909**	36.93	0.969	30.95	0.892	33.04	0.922	33.34	0.936	32.80	0.925
HINet [7]	30.29	0.905	37.28	0.970	30.65	0.894	33.05	0.919	33.91	0.940	33.03	0.926
IDLIR [36]	28.33	0.894	35.72	0.965	29.33	0.886	32.06	0.917	32.93	0.936	31.67	0.920
Uformer-B [51]	28.71	0.896	35.91	0.964	27.54	0.871	32.34	0.913	30.88	0.928	31.08	0.914
IDT [54]	29.69	0.905	37.01	0.971	29.95	0.898	31.38	0.908	33.38	0.937	32.28	0.924
Semi-Swin [40]	28.54	0.893	34.71	0.957	28.79	0.861	30.96	0.909	32.68	0.932	31.14	0.910
Restormer [64]	30.86	0.906	37.56	0.974	**31.46**	**0.904**	33.19	**0.926**	33.98	**0.942**	33.41	0.930
PSTNet	**31.16**	0.905	**39.52**	**0.980**	31.21	0.903	**33.35**	0.925	**34.20**	0.941	**33.88**	**0.931**

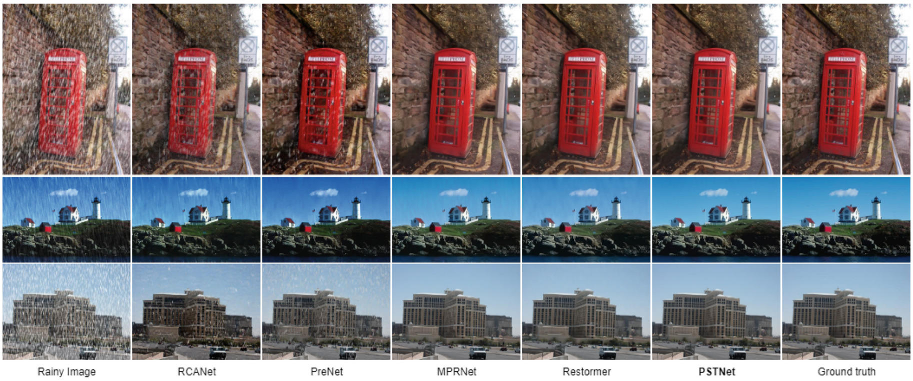


Fig. 2. Our PSTNet produces rain-free images that retain structural integrity.

4.1 Ablation Studies

Effectiveness of Selecting Highest K Attentions. To assess the contribution of selective attention in the SMHA, we compare the deraining results of SMHA without selective attention (see Table 4). The PSNR values of images processed by selecting the highest k attention values are better than those without them. Our method reconstructs finer features and enhances the restoration quality compared to normal self-attention operations without selective attention. Long-range pixel dependencies are less likely to contain unnecessary context

Table 3. Model computational complexity evaluation for an input size 256 × 256

Method	Params (M)	MACs (G)
MPRNet	20.1	778.2
HINet	88.7	170.7
Restormer	26.13	140
PSTNet	12.4	89.1

Table 4. Ablation experiments for highest-k selection. PSNR is computed for the datasets.

Dataset	w/o highest-k	w highest-k
Rain100L	39.32	39.52
Rain100H	31.13	31.21

when using the highest-k selection operator since neighbouring pixels are more comparable than distant ones. During the self-attention computation, this selection phase eliminates smaller similarity weights from certain long-range feature interactions, improving representation accuracy and producing higher-quality output.

Effect of k Value. The effect of the key parameter k of our proposed SMHA is analysed in Table 5. The k value significantly influences the sparsity. If value of k fixed, like $\frac{1}{2}$, it greatly affects how well it performs. We construct a configurable interval range for k to prevent an exhaustive search. The model dynamically determines the most contributive score. Performance suffers greatly when k is too small because insufficient global information is captured. The best result, 31.21 dB, is achieved when $[\Delta_1, \Delta_2]$ for SMHA is set in the range $[\frac{1}{2}, \frac{4}{5}]$. Nevertheless, performance declines as k increases because of the addition of irrelevant characteristics.

Table 5. Ablation experiments for k value in SMHA. PSNR is computed for Rain100H

k	$[\frac{1}{5}, \frac{1}{2}]$	$[\frac{1}{4}, \frac{2}{3}]$	$[\frac{1}{2}, \frac{4}{5}]$	$[\frac{2}{3}, \frac{5}{6}]$
PSNR	30.60	31.12	31.21	31.16

Effectiveness of SGFN. We evaluate the suggested SGFN by contrasting it with GDFN [64]. Compared to the complex implementation of GELU, simpleGate is easy to implement. By replacing GELU of GDFN with SimpleGate, the image deraining performance (PSNR) (on Test100) is increased from 31.09 to 31.16. The results show that SimpleGate can replace GELU.

5 Conclusion

We developed an efficient sparse Transformer network, PSTNet, for image deraining. Significant improvements are made to the sparse Transformer's primary components to improve feature aggregation and transformation. Observing that vanilla self-attention in Transformers can be hampered by global interactions with irrelevant information. we developed selective attention, which keeps the most valuable self-attention values. The proposed simpleGate feed-forward network (SGFN) also simplifies the gating mechanism for controlled feature transformation. Experimental results demonstrate that our PSTNet performs better than state-of-the-art methods.

Acknowledgements. The authors express their gratitude to the Council of Science and Technology, U.P. (CST-UP), India, for their financial assistance in performing this research work. This work is one of the project's outcomes entitled "Road object detection during rainy nights for autonomous vehicles" with sanction no. CST/D-1396 dated 20.10.2023, sponsored by CST-UP, India.

References

1. Abuolaim, A., Brown, M.S.: Defocus deblurring using dual-pixel data. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J.-M. (eds.) ECCV 2020, Part X. LNCS, vol. 12355, pp. 111–126. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-58607-2_7
2. Ba, J.L., Kiros, J.R., Hinton, G.E.: Layer normalization. arXiv preprint arXiv:1607.06450 (2016)
3. Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., Zagoruyko, S.: End-to-end object detection with transformers. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J.-M. (eds.) ECCV 2020. LNCS, vol. 12346, pp. 213–229. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-58452-8_13
4. Chen, D., et al.: Gated context aggregation network for image dehazing and deraining. In: WACV 2019 (2018)
5. Chen, H., et al.: Pre-trained image processing transformer. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 12299–12310 (2021)
6. Chen, L., Chu, X., Zhang, X., Sun, J.: Simple baselines for image restoration. In: Avidan, S., Brostow, G., Cissé, M., Farinella, G.M., Hassner, T. (eds.) ECCV 2022. LNCS, vol. 13667, pp. 17–33. Springer, Cham (2022). https://doi.org/10.1007/978-3-031-20071-7_2
7. Chen, L., Lu, X., Zhang, J., Chu, X., Chen, C.: Hinet: half instance normalization network for image restoration. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 182–192 (2021)
8. Chen, T., Cheng, Y., Gan, Z., Yuan, L., Zhang, L., Wang, Z.: Chasing sparsity in vision transformers: An end-to-end exploration. In: Advances in Neural Information Processing Systems, vol. 34, pp. 19974–19988 (2021)
9. Chen, X., Li, H., Li, M., Pan, J.: Learning a sparse transformer network for effective image deraining. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 5896–5905 (2023)

10. Chen, X., Pan, J., Lu, J., Fan, Z., Li, H.: Hybrid CNN-transformer feature fusion for single image deraining. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 37, pp. 378–386 (2023)
11. Cho, S.J., Ji, S.W., Hong, J.P., Jung, S.W., Ko, S.J.: Rethinking coarse-to-fine approach in single image deblurring. In: IEEE/CVF International Conference on Computer Vision, pp. 4641–4650 (2021)
12. Correia, G.M., Niculae, V., Martins, A.F.: Adaptively sparse transformers. arXiv preprint arXiv:1909.00015 (2019)
13. Dosovitskiy, A., et al.: An image is worth 16x16 words: transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
14. Fu, X., Huang, J., Ding, X., Liao, Y., Paisley, J.: Clearing the skies: a deep network architecture for single-image rain removal. IEEE Trans. Image Process. **26**(6), 2944–2956 (2017)
15. Fu, X., Huang, J., Zeng, D., Huang, Y., Ding, X., Paisley, J.: Removing rain from single images via a deep detail network. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3855–3863 (2017)
16. Fu, X., Xiao, J., Zhu, Y., Liu, A., Wu, F., Zha, Z.J.: Continual image deraining with hypergraph convolutional networks. IEEE Trans. Pattern Anal. Mach. Intell. **45**(8), 9534–9551 (2023)
17. Fu, Z., Fu, Z., Liu, Q., Cai, W., Wang, Y.: Sparsett: visual tracking with sparse transformers. arXiv preprint arXiv:2205.03776 (2022)
18. Gu, S., Li, Y., Gool, L.V., Timofte, R.: Self-guided network for fast image denoising. In: IEEE/CVF International Conference on Computer Vision, pp. 2511–2520 (2019)
19. Gu, S., Meng, D., Zuo, W., Zhang, L.: Joint convolutional analysis and synthesis sparse representation for single image layer separation. In: IEEE International Conference on Computer Vision, pp. 1708–1716 (2017)
20. Hu, X., Fu, C.W., Zhu, L., Heng, P.A.: Depth-attentional features for single-image rain removal. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 8022–8031 (2019)
21. Huang, H., Yu, A., He, R.: Memory oriented transfer learning for semi-supervised image deraining. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 7732–7741 (2021)
22. Huynh-Thu, Q., Ghanbari, M.: Scope of validity of PSNR in image/video quality assessment. Electron. Lett. **44**(13), 800–801 (2008)
23. Jiang, K., Wang, Z., Chen, C., Wang, Z., Cui, L., Lin, C.W.: Magic ELF: image deraining meets association learning and transformer. arXiv preprint arXiv:2207.10455 (2022)
24. Jiang, K., et al.: Multi-scale progressive fusion network for single image deraining. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 8346–8355 (2020)
25. Kang, L.W., Lin, C.W., Fu, Y.H.: Automatic single-image-based rain streaks removal via image decomposition. IEEE Trans. Image Process. **21**(4), 1742–1755 (2011)
26. Khan, S., Naseer, M., Hayat, M., Zamir, S.W., Khan, F.S., Shah, M.: Transformers in vision: a survey. ACM Comput. Surv. (CSUR) **54**(10s), 1–41 (2022)
27. Kupyn, O., Martyniuk, T., Wu, J., Wang, Z.: Deblurgan-v2: deblurring (orders-of-magnitude) faster and better. In: IEEE/CVF International Conference on Computer Vision, pp. 8878–8887 (2019)

28. Li, X., Wu, J., Lin, Z., Liu, H., Zha, H.: Recurrent squeeze-and-excitation context aggregation net for single image deraining. In: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (eds.) ECCV 2018. LNCS, vol. 11211, pp. 262–277. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-01234-2_16
29. Li, Y., Tan, R.T., Guo, X., Lu, J., Brown, M.S.: Rain streak removal using layer priors. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 2736–2744 (2016)
30. Liang, Y., Anwar, S., Liu, Y.: DRT: a lightweight single image deraining recursive transformer. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 589–598 (2022)
31. Liu, X., Suganuma, M., Sun, Z., Okatani, T.: Dual residual networks leveraging the potential of paired operations for image restoration. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 7007–7016 (2019)
32. Liu, Y., Yue, Z., Pan, J., Su, Z.: Unpaired learning for deep image deraining with rain direction regularizer. In: IEEE/CVF International Conference on Computer Vision, pp. 4753–4761 (2021)
33. Liu, Z., et al.: Swin transformer: hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 10012–10022 (2021)
34. Loshchilov, I., Hutter, F.: SGDR: stochastic gradient descent with warm restarts. arXiv preprint arXiv:1608.03983 (2016)
35. Luo, Y., Xu, Y., Ji, H.: Removing rain from a single image via discriminative sparse coding. In: IEEE International Conference on Computer Vision, pp. 3397–3405 (2015)
36. Ma, M., Ren, D., Yang, Y.: Integrating degradation learning into image restoration. In: IEEE International Conference on Multimedia and Expo, pp. 1–6 (2022). https://doi.org/10.1109/ICME52920.2022.9859813
37. 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)
38. Purohit, K., Suin, M., Rajagopalan, A., Boddeti, V.N.: Spatially-adaptive image restoration using distortion-guided networks. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 2309–2319 (2021)
39. Qin, Q., Yan, J., Wang, Q., Wang, X., Li, M., Wang, Y.: ETDNET: an efficient transformer deraining model. IEEE Access **9**, 119881–119893 (2021)
40. Ren, C., Yan, D., Cai, Y., Li, Y.: Semi-swinderain: semi-supervised image deraining network using swin transformer. In: IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 1–5 (2023). https://doi.org/10.1109/ICASSP49357.2023.10095214
41. Ren, D., Zuo, W., Hu, Q., Zhu, P., Meng, D.: Progressive image deraining networks: a better and simpler baseline. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 3937–3946 (2019)
42. 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,Part III. LNCS, vol. 9351, pp. 234–241. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-24574-4_28
43. Roy, A., Saffar, M., Vaswani, A., Grangier, D.: Efficient content-based sparse attention with routing transformers. Trans. Assoc. Comput. Linguist. **9**, 53–68 (2021)
44. Shi, W., et al.: 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)

45. Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A., Jégou, H.: Training data-efficient image transformers & distillation through attention. In: International Conference on Machine Learning, pp. 10347–10357. PMLR (2021)
46. Vaswani, A., et al.: Attention is all you need. In: Advances in Neural Information Processing Systems, vol. 30 (2017)
47. Wang, H., Shen, J., Liu, Y., Gao, Y., Gavves, E.: Nformer: robust person re-identification with neighbor transformer. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 7297–7307 (2022)
48. Wang, P., et al.: KVT: k-NN attention for boosting vision transformers. In: Avidan, S., Brostow, G., Cissé, M., Farinella, G.M., Hassner, T. (eds.) ECCV 2022. LNCS, vol. 13684, pp. 285–302. Springer, Cham (2022). https://doi.org/10.1007/978-3-031-20053-3_17
49. Wang, W., et al.: 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)
50. Wang, Y., Ma, C., Zeng, B.: Multi-decoding deraining network and quasi-sparsity based training. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 13375–13384 (2021)
51. Wang, Z., Cun, X., Bao, J., Zhou, W., Liu, J., Li, H.: Uformer: a general u-shaped transformer for image restoration. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 17683–17693 (2022)
52. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: from error visibility to structural similarity. IEEE Trans. Image Process. **13**(4), 600–612 (2004)
53. Wei, W., Meng, D., Zhao, Q., Xu, Z., Wu, Y.: Semi-supervised transfer learning for image rain removal. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 3877–3886 (2019)
54. Xiao, J., Fu, X., Liu, A., Wu, F., Zha, Z.J.: Image de-raining transformer. IEEE Trans. Pattern Anal. Mach. Intell. (2022)
55. Xie, E., Wang, W., Yu, Z., Anandkumar, A., Alvarez, J.M., Luo, P.: SegFormer: simple and efficient design for semantic segmentation with transformers. In: Advances in Neural Information Processing Systems, vol. 34, pp. 12077–12090 (2021)
56. Yang, W., Tan, R.T., Feng, J., Liu, J., Guo, Z., Yan, S.: Deep joint rain detection and removal from a single image. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1357–1366 (2017)
57. Yang, W., Tan, R.T., Wang, S., Fang, Y., Liu, J.: Single image deraining: from model-based to data-driven and beyond. IEEE Trans. Pattern Anal. Mach. Intell. **43**(11), 4059–4077 (2020)
58. Yasarla, R., Patel, V.M.: Uncertainty guided multi-scale residual learning-using a cycle spinning CNN for single image de-raining. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 8405–8414 (2019)
59. Yasarla, R., Sindagi, V.A., Patel, V.M.: Syn2real transfer learning for image deraining using gaussian processes. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 2726–2736 (2020)
60. Ye, Y., Chang, Y., Zhou, H., Yan, L.: Closing the loop: joint rain generation and removal via disentangled image translation. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 2053–2062 (2021)
61. Yuan, K., Guo, S., Liu, Z., Zhou, A., Yu, F., Wu, W.: Incorporating convolution designs into visual transformers. In: IEEE/CVF International Conference on Computer Vision, pp. 579–588 (2021)

62. Yuan, L., et al.: 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)
63. Yue, Z., Zhao, Q., Zhang, L., Meng, D.: Dual adversarial network: toward real-world noise removal and noise generation. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J.-M. (eds.) ECCV 2020, Part X. LNCS, vol. 12355, pp. 41–58. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-58607-2_3
64. 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)
65. Zamir, S.W., et al.: Learning enriched features for real image restoration and enhancement. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J.-M. (eds.) ECCV 2020, Part XXV. LNCS, vol. 12370, pp. 492–511. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-58595-2_30
66. Zamir, S.W., et al.: Multi-stage progressive image restoration. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 14821–14831 (2021)
67. Zhang, H., Patel, V.M.: Convolutional sparse and low-rank coding-based rain streak removal. In: IEEE Winter Conference on Applications of Computer Vision, pp. 1259–1267. IEEE (2017)
68. Zhang, H., Patel, V.M.: Density-aware single image de-raining using a multi-stream dense network. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 695–704 (2018)
69. Zhang, H., Sindagi, V., Patel, V.M.: Image de-raining using a conditional generative adversarial network. IEEE Trans. Circuits Syst. Video Technol. **30**(11), 3943–3956 (2019)
70. Zhang, J., Zhang, Y., Gu, J., Zhang, Y., Kong, L., Yuan, X.: Accurate image restoration with attention retractable transformer. arXiv preprint arXiv:2210.01427 (2022)
71. Zhang, K., Li, Y., Zuo, W., Zhang, L., Van Gool, L., Timofte, R.: Plug-and-play image restoration with deep denoiser prior. IEEE Trans. Pattern Anal. Mach. Intell. **44**(10), 6360–6376 (2021)
72. Zhang, Y., Li, K., Li, K., Zhong, B., Fu, Y.: Residual non-local attention networks for image restoration. arXiv preprint arXiv:1903.10082 (2019)
73. Zhao, G., Lin, J., Zhang, Z., Ren, X., Su, Q., Sun, X.: Explicit sparse transformer: concentrated attention through explicit selection. arXiv preprint arXiv:1912.11637 (2019)
74. Zheng, S., et al.: Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 6881–6890 (2021)
75. Zhu, X., Su, W., Lu, L., Li, B., Wang, X., Dai, J.: Deformable DeTR: deformable transformers for end-to-end object detection. arXiv preprint arXiv:2010.04159 (2020)

Frequency Modulated Deformable Transformer for Underwater Image Enhancement

Adinath Dukre[1](✉), Vivek Deshmukh[1], Ashutosh Kulkarni[2], Shruti Phutke[3], Santosh Kumar Vipparthi[2], Anil B. Gonde[1], and Subrahmanyam Murala[4]

[1] Shri Guru Gobind Singhji Institute of Engineering and Technology, Nanded, India
2020bec100@sggs.ac.in

[2] CVPR Lab, Indian Institute of Technology Ropar, Rupnagar, India

[3] ETI Lab, Yamaha Motor Solutions, Faridabad, India

[4] School of Computer Science and Statistics, Trinity College Dublin, Dublin, Ireland

Abstract. Underwater images frequently experience quality degradation due to refraction, back-scattering, and absorption, leading to color distortion, blurriness, and reduced visibility. Such degradation present in the underwater images can cause inaccuracies while functioning with higher advanced level computer vision applications, equipped for autonomous underwater vehicles. Despite the ability of enhancing the degraded images, existing approaches fail at preserving the localized fine edges also producing the true colors. Therefore, an effective preprocessing network is necessary for underwater image enhancement. With this motivation, we propose a frequency modulated deformable transformer network for underwater image enhancement. Initially, the features are extracted with the proposed multi-scale feature fusion feed-forward module. Further, the frequency modulated deformable attention module is proposed to reconstruct fine-level texture in the restored image. Here, we propose a spatio-channel attentive offset extractor in the modulated deformable convolution for focusing on relevant contextual information. Also, adaptive edge-preserving skip connections are proposed for propagating prominent edge features from the network's shallow layers to its deeper layers. A comprehensive evaluation of the proposed method on synthetic and real-world datasets and extensive ablation analysis demonstrates that the proposed approach shows superior performance than existing state-of-the-art methods. The testing code is provided at https://github.com/adinathdukre/FMDTUIE.

Keywords: Deformable Transformer · Spatio-Channel Attentive · Offset Extraction · Underwater Image Enhancement

A. Dukre and V. Deshmukh—Equal contributions were made by these authors to the work.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/978-3-031-78125-4_9.

A. Antonacopoulos et al. (Eds.): ICPR 2024, LNCS 15332, pp. 121–136, 2025.
https://doi.org/10.1007/978-3-031-78125-4_9

1 Introduction

Recent advancements have seen the use of autonomous underwater vehicles (AUVs) furnished with visual perception systems to collect high-quality photographs in hazardous and contaminated environments, such as underwater archaeology, marine environment monitoring, *etc.*. Additionally, sub-tasks like object grasping, object recognition, and image segmentation, *etc.,* are routinely involved in these applications to maintain efficient performance. These sub-tasks need clean data as input. However, capturing high-quality clean underwater photos is challenging due to the wavelength dependent absorption, reflection, and scattering issues, leading to hazy blur, color cast, and restricted visibility. Therefore, an effective underwater image enhancement (UIE) is highly favorable to maintain the significant performance of these sub-tasks.

Although remarkable success has been achieved in underwater image enhancement, the problem of restoring proper textural detail in an image is still an open challenge. Existing works such as histogram distribution [13], prior probability [7], and attenuation prior [27] are not adaptive to the varied underwater degraded circumstances. Also, researchers introduced various deep learning-based methods for achieving adaptation to the varying degradations present in underwater images. However, These techniques utilize a convolution operator with a restricted receptive field, which limits their ability to capture long-range pixel dependencies. To handle this, the UIE network should have a significantly adaptive receptive field for capturing the long-range pixel dependencies.

To address the above problem, researchers have proposed vision transformer [29] based approaches due to its capabilities of capturing long-range dependencies for UIE. Shen *et al.* [23] proposed a transformer with depth-wise convolution

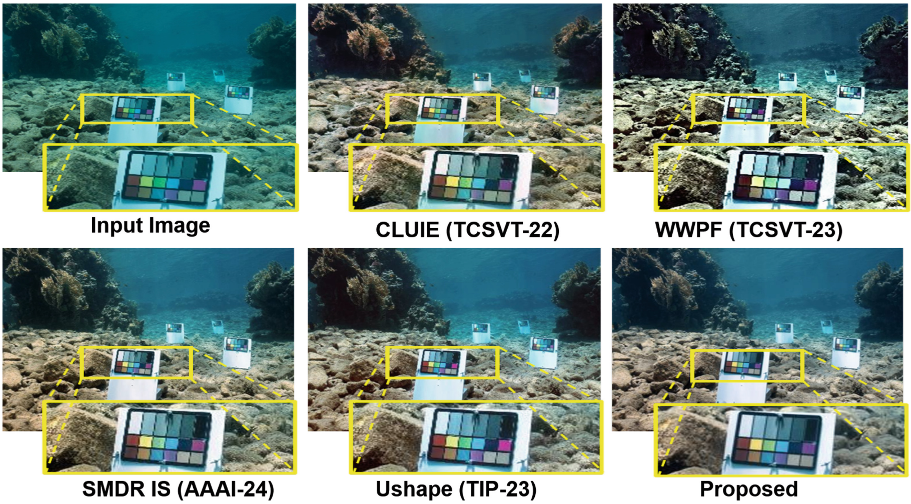


Fig. 1. Sample visual results of the proposed and existing state-of-the-art methods. *Existing methods are unable to generate the localized edges whereas the proposed approach is able to generate effective localized edges and true color.* (Color figure online)

and multi-head self-attention to extract low-level features and capture structure variations of the object. In [19], dilated convolution is employed to make the network lightweight and expand the receptive field. Tang *et al.* [25] proposed a neural architecture search-based transformer approach for UIE. Another transformer-based approach along with gray-scale attention is proposed in [11]. Even though these approaches produce fruitful results, structural details in the enhanced images are missing. Further, [23,26] approaches employ transformer architecture with direct skip connections between encoder-decoder, which may transfer the degraded information from the shallow layer to the deep layer. Shen *et al.* [23], and Wang *et al.* [26] employed a feed-forward network from a vision transformer for UIE. Further, Liu *et al.* [17] fused two scale features in a feed-forward network. As a result, these methods struggle to capture and reconstruct images with more *localized edges and textural details.*

Motivated by the above challenge, we propose frequency modulated deformable transformer network for underwater image enhancement. In order to capture the structural variations in the input image, we propose a frequency modulated deformable attention (FMDA) module. Also, we propose the adaptive edge-preserving (AEP) module to traverse the fine edges without degradation from shallow layers to deep layers via skip connections. Further, we propose the multi-scale feature fusion (MSF) module to capture more localized edges and textural features during restoration. Our main contributions are:

- We propose a frequency modulated deformable attention module with multi-scale feature fusion-based feed-forward architecture for underwater image enhancement.
- Spatio-channel attentive offset extractor is proposed in modulated deformable convolution for extracting color correlation and spatially relevant information from the features.
- The adaptive edge-preserving module is proposed to forward the structural information from the encoder to the respective decoder for effective enhancement.

Comprehensive experimental study, on synthetic and real-world datasets depicts our proposed underwater image enhancement method is superior to existing methods. The sample visual results analysis among proposed and previous methods is provided in Fig. 1, which shows that the proposed method preserves the *structural details with fine edges* and *true colors* in the image along with minimizing degradations.

2 Related Work

Over the recent years, numerous techniques for restoring and enhancing underwater images have emerged, aiming to elevate the visual excellence of such images.

2.1 Traditional Methods

Earlier research on underwater image enhancement relied on handcrafted and model-based approaches. Hitam *et al.* [8] employed contrast manipulation and adaptive histogram equalization techniques in both RGB and HSV color spaces to augment the contrast of underwater photographs and decrease noise levels. Fu *et al.* [7] introduced the retinex model which encompasses layer decomposition, color correction, and enhancement techniques for UIE. Similar to air-medium dark-channel prior that modifies preceding dark-channel prior is proposed in [4]. Moreover, Huang *et al.* [9] presented a technique that employs dynamic hyperparameter-based histogram stretching as well as bilateral filters to preserve details for underwater image enhancement. However, these methods mainly rely on assumptions on which priors are defined and fail to cope with real complex scenarios.

2.2 Deep Learning-Based Methods

In recent years, the use of deep learning techniques has become more significant in addressing issues in computer vision. For UIE, Islam *et al.* [10] and Fabbri *et al.* [5] employed conditional generative adversarial network (CGAN). In [14], authors introduced underwater image enhancement convolutional neural network (UWCNN). However, this method employs a convolutional operator which has a restricted receptive field. Therefore, it does not account for fine structural details of the image. Sharma *et al.* [22] suggested an attention-based and multi-receptive network that performs both underwater image enhancement and super-resolution simultaneously. Further, the color, global, and local contrast issues are solved in [31]. Li *et al.* [17] proposed a color histogram approach for UIE. These methods achieve superior performance but focus only on maintaining color details. However, enhancing the structural information has equal importance, which is ignored in the above-discussed methods.

2.3 Transformers for Image Restoration

The transformer architecture leverages self-attention, where attention coefficients signify the interplay between data on both global and local scales [6]. Therefore, transformers are extensively employed for diverse image restoration tasks [29]. Zamir *et al.* [29] developed an effective transformer network that can be used for a variety of restoration tasks, including image de-raining, and deblurring. Liu *et al.* [20] introduced the "Swin Transformer" which calculates attention within shifted windows to reduce the computational load in tasks like image de-noising and de-blurring. Tang *et al.* [25] proposed a neural architecture search-based transformer approach. Wang *et al.* [26] proposed a network that takes swin transformer block as its basic unit for UIE. Further, these methods employ the simple skip connection which may traverse the degradation from shallow to deep features.

However, considering the above issues, we propose a frequency modulated deformable transformer UIE network which preserves structural detail and color along with reducing degradation's.

3 Proposed Framework

The schematic of the proposed network for UIE is provided in Fig. 2. The detailed significance of each proposed module is provided in the next subsections.

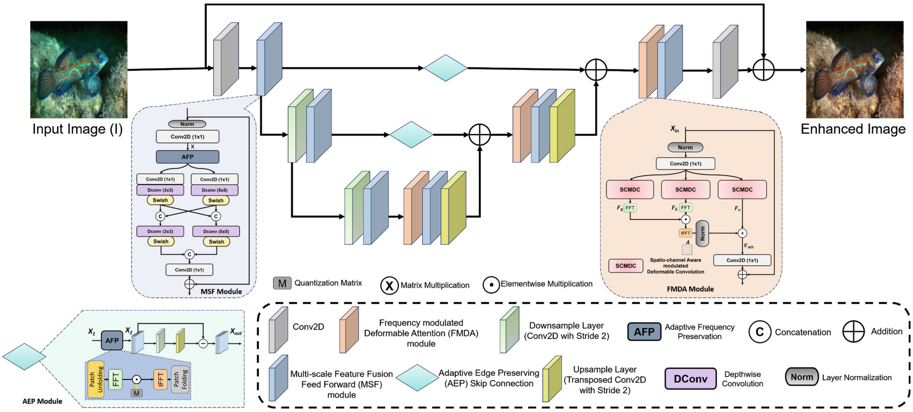


Fig. 2. An architectural diagram of proposed method for enhancing underwater images.

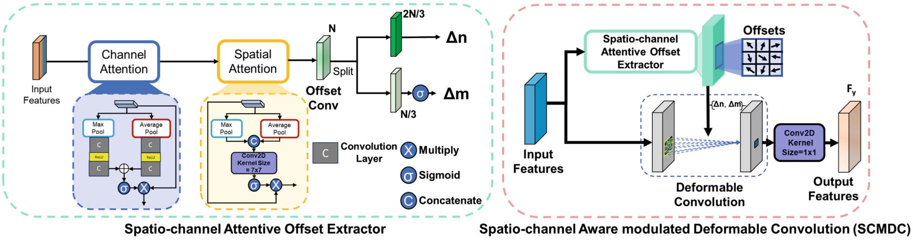


Fig. 3. Illustration of proposed spatio-channel attentive offset extractor and spatio-channel aware deformable convolution.

3.1 Multi-scale Feature Fusion Feed-Forward (MSF) Module

Existing feed-forward module based transformer network [23] are unable to process high-frequency components like texture, edge information, *etc.*. Also, they are not capable of capturing more fine details and contextual information. To handle this issue, we propose multi-scale feature fusion feed-forward (MSF) module. In proposed MSF module (*refer MSF module from* Fig. 2), we first exploit the adaptive frequency preserving (AFP) block based on the reverse process of the JPEG compression algorithm which contains quantization learnable matrix (*for more details refer* Sect. 3.3). In MSF, the input tensor X is given to the AFP block after applying layer normalization and 1×1 convolution. Here, quantization learnable matrix in AFP block is used to learn processing of the high-frequency components and restrict the low frequencies like hazy blur [12]. Further, the output of the AFP is given to two parallel different paths that utilize 1×1 convolution Succeeded by a depth-wise convolution given kernel size 5×5, and 3×3 with Swish activation function. Here, we have integrated two multi-scale depth-wise convoluted features with each other and passed through respective depth-wise convolution followed by Swish activation function. Finally, these features are merged to capture fine details and improve the local and contextual information.

3.2 Frequency Modulated Deformable Attention (FMDA) Module

Transformers are adept at capturing long-range dependencies using self-attention, their superiority over conventional CNNs and GANs on both high-level and low-level vision tasks like segmentation, object detection, deblurring, dehazing, deraining, and denoising, *etc.,* is remarkable. Also, modulated deformable convolutions have proven to be more effective due to their ability to accommodate the shape variation of objects. However, the attention with depth-wise convolution may suffer from limited receptive fields [28], which restricts the overall network from capturing structural variations present in the image. To tackle this issue, we proposed spatio-channel aware modulated deformable convolution (SCMDC) for extracting features of queries (F_q), keys (F_k), and values (F_v) as:

$$F_q, F_k, F_v = SCMDC_{3\times3}\left(C_1(L(X_{in}))\right) \tag{1}$$

where, $SCMDC_{3\times3}(\cdot)$ is spatio-channel aware modulated deformable convolution (*see Fig.* 3), C_1 is convolution with 1×1 kernel, and $L(\cdot)$ is layer normalization.

The offsets in modulated deformable convolution may exceed their contextually relevant regions [32], resulting in the emergence of irrelevant features and, the formation of partially restored pictures. To address this, we have introduced a spatio-channel attentive offset extractor that is sensitive to color shifts induced by underwater conditions (*see* Fig. 3). Here, the extraction of offsets and modulation values originates from the same offset convolution process, employing channel-wise spatially attentive features as its input as:

$$F_y = \sum_{i=1}^{N} DFconv_{3\times3}\left(X_{n+n_i+\Delta n_i}\right)\Delta m_i \tag{2}$$

where, N represents a sampling location within a 3×3 convolutional grid, $DFconv_{3\times3}(\cdot)$ denotes a modulated deformable convolution with a 3×3 kernel size and $y \in (q,k,v)$. The variable n signifies a feature location, while Δn represents the offsets obtained from the spatio-channel attentive offset extractor.

Similarly, Δm denotes the extracted modulator scalars from the Spatio-Channel attentive offset extractor block, and $n_i \in \{(-1, \text{-}1), (-1, 0)\ldots(1, 1)\}$. Feature map visualization of various combinations of offset extractors (*see* Fig. 4) shows proposed *SCMDC* offset extractor can extract more local spatial information. With this process, we have extracted the F_q, F_v, and F_k *(More detailed information is available in the supplementary material.)*.

Further, to reduce overall computation cost, the frequency domain correlation between F_q and F_k is calculated (*see FMDA in* Fig. 2) with fast Fourier transform (FFT) [12] as:

$$A = F^{-1}\left(F\left(F_q\right).\overline{F\left(F_k\right)}\right) \tag{3}$$

where, $F(\cdot)$ represents the FFT, $F^{-1}(\cdot)$ represents the inverse FFT, and $\overline{F(\cdot)}$ represents the conjugate transpose operation. Lastly, we compute the summarized feature through:

$$V_{att} = L(A).F_v \tag{4}$$

where, $L(\cdot)$ is layer normalization. Finally, the output features of FMDA are generated as:

$$FMDA = X_{in} + Conv_{1\times 1}\left(V_{att}\right) \tag{5}$$

where, X_{in} is the input features. Spatio-channel aware modulated deformable convolution-based frequency domain self-attention layer, likewise known as the frequency modulated deformable attention module ($FMDA$). This proposed FMDA module used three times at various levels to get the enhanced image.

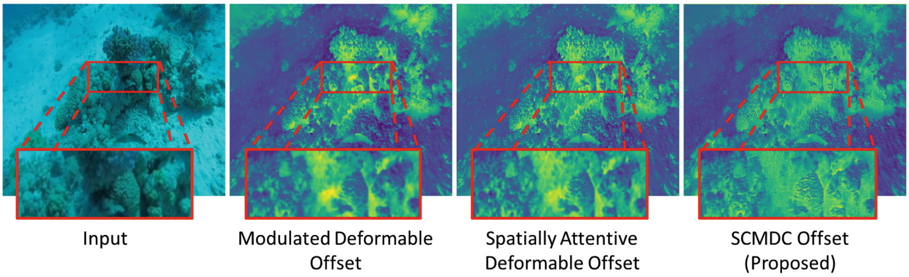

Input
Modulated Deformable Offset
Spatially Attentive Deformable Offset
SCMDC Offset (Proposed)

Fig. 4. Feature map visualization of various combinations of offset extractor. The proposed SCMDC offset extractor can extract more local spatial information (*as shown in the red box*) than the modulated deformable offset and spatially attentive deformable offset extractor, resulting in a superior structural variation in the proposed method output images. (Color figure online)

3.3 Adaptive Edge Preserving (AEP) Module

To achieve better performance in our UIE task, we must deal with preserving edge-sensitive regions, and reduce irrelevant information propagated via feature extraction. The previous approaches [23,26] employed direct skip connections that passes the extracted features without refining and considering sensitive information like edges. To tackle this issue, we proposed an adaptive edge-preserving module for processing the features during skip connection.

Initially, the input features X_1 are passed through adaptive frequency preservation (AFP) block [12] to generate refined the features X_2. The details of AFP block are shown in Fig. 2. The output of AFP (X_2) is given as:

$$X_2 = P^{-1}\left(F^{-1}\left(X_1^f\right)\right); X_1^f = F\left(P\left(X_1\right)\right) \tag{6}$$

where, $P(\cdot)$ and $P^{-1}(\cdot)$ represents patch unfolding and folding operations respectively, $F(\cdot)$ and $F^{-1}(\cdot)$ denotes FFT and the inverse FFT . After FFT, on the transformed features, *the learnable quantization matrix is used to process high frequency information and suppress low frequency component present in the feature maps.*

Further, these refined features from AFP are passed through downsample-upsample operator and subtracted from the refined input feature map to pass only refined edge and texture information as:

$$X_{out} = C_1(X_2 - D_2(C_2(X_2))) \tag{7}$$

where, D_2 and C_2 represents de-convolution with up-sampling factor 2 and convolution with down-sampling factor 2 respectively (*see adaptive edge preserving skip connection in* Fig. 2). Overall the proposed adaptive edge preserving module assist the proposed network by obtaining refined high-frequency edge information. The Performance of all proposed blocks is examined in the ablation studies (refer Sect. 5.4 for more details).

4 Training Details

We trained the proposed network on the EUVP [10] and UIEB [15] dataset. During training phase, we crop randomly the original input image into a 256 × 256 patch Size. Also, data augmentation like horizontal flip and vertical flip is adopted to make enough training samples. The patch size for the quantization matrix estimation in the adaptive edge-preserving module is empirically set to be 8×8 [12]. To optimize ours proposed network parameters, we have used ADAM as an optimizer with initial learning rate of the 10^{-4} and a minimum of 10^{-7}, which is changed with the cosine annealing technique. The network is trained on Nvidia Titan Xp having a 2.2 GHz clock speed. The various losses like $\mathbb{L}_1$, FFT, perceptual, and contrastive are employed to optimize the performance of proposed network. The details of each loss function are given: **Loss Functions:** The network is trained with the content loss function (L_1). Further to reduce difference in between frequency space, we have employed FFT loss ($\mathbb{L}_F$) [3] which calculates the likeness between ground truth and network output. Furthermore to maintain the feature level textural and structure similarity, the perceptual loss (L_p) is computed using the VGG-16 [24] pre-trained module. Also, the contrastive loss ($\mathbb{L}_C$) is calculated to maximize and minimize the difference between input-output and output-ground truth respectively Therefore, the total loss is represented as:

$$\mathbb{L}_{total} = \lambda_1 \mathbb{L}_1 + \lambda_2 \mathbb{L}_F + \lambda_3 \mathbb{L}_p + \lambda_4 \mathbb{L}_C \tag{8}$$

we set weights as $\lambda_1 = 1, \lambda_2 = 0.1, \lambda_3 = 15$ and $\lambda_4 = 5$. *The detailed explanation and ablation for loss functions is available in supplementary material.*

5 Experimental Analysis

5.1 Datasets

EUVP [10]: The Enhancing Underwater Visual Perception (EUVP) database covers underwater images captured with many types of cameras in various configurations. The dataset comprises 11,435 image pairs (clean and degraded) for training and 515 image pairs for testing.

UIEB [15]: The underwater image enhancement benchmark (UIEB) dataset contains 890 underwater images from various scenarios. From this, the training set is created by randomly picking 800 images and the remaining 90 images for testing.

Sea-Thru [1]: We used 10 real-world underwater images from this dataset for qualitative and non-reference evaluation.

Color-Checker [2]: This dataset consists of 7 real-world underwater images, we have utilized this dataset to evaluate color correction effectiveness based on qualitative and non-reference color metrics.

Table 1. Analysis of the proposed (Ours) and existing methods on the UIEB [15] and EUVP [10] dataset in the terms of an average SSIM (↑), PSNR (↑), and LPIPS (↓) for underwater image enhancement (Note: ↓: *lower is better*, ↑: *higher is better*).

Method	Publication	UIEB			EUVP		
		PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓
LANet [18]	RAL-22	24.05	0.90	0.13	25.82	0.86	0.28
CLUIE [16]	TCSVT-22	20.37	0.89	0.18	-	-	-
Wave Net [22]	TMCCA-23	21.57	0.80	0.12	28.62	0.83	0.24
WWPF [31]	TCSVT-23	18.59	0.79	0.22	-	-	
U-shape [21]	TIP-23	21.39	0.85	0.24	26.77	0.87	0.26
SMDR-IS [30]	AAAI-24	23.71	0.92	0.14	-	-	-
Ours	-	**25.79**	**0.95**	**0.11**	**30.90**	**0.92**	**0.22**

Table 2. Evaluations of different methods on non-reference metrics for Color-checker [2] and Sea-thru [1] dataset in terms of UIQM (↑), UCIQE (↑), UICM (↑) and NIQE (↓) (Note: ↓: *lower is better*, ↑: *higher is better*).

Method	Publication	Color-checker				Sea-thru			
		UIQM	UCIQE	UICM	NIQE	UIQM	UCIQE	UICM	NIQE
LANet [18]	RAL-22	4.06	32.89	−24.59	**3.09**	4.57	29.54	−19.74	4.39
WaveNet [22]	TMCCA-23	4.13	33.15	−21.02	3.14	4.60	30.25	−21.24	5.22
CLUIE [16]	TCSVT-22	4.43	33.58	−18.64	3.10	3.91	29.86	−37.72	4.61
WWPF [31]	TCSVT-23	3.96	33.03	−26.86	3.10	3.84	30.50	−18.00	6.43
U-shape [21]	TIP-23	4.03	30.65	−13.91	5.30	4.01	29.05	−22.55	4.54
SMDR-IS [30]	AAAI-24	3.97	32.84	−31.76	3.11	4.20	30.35	−28.48	4.86
Ours	-	**4.53**	**33.67**	−**7.84**	3.17	**4.74**	30.29	−**10.23**	**3.57**

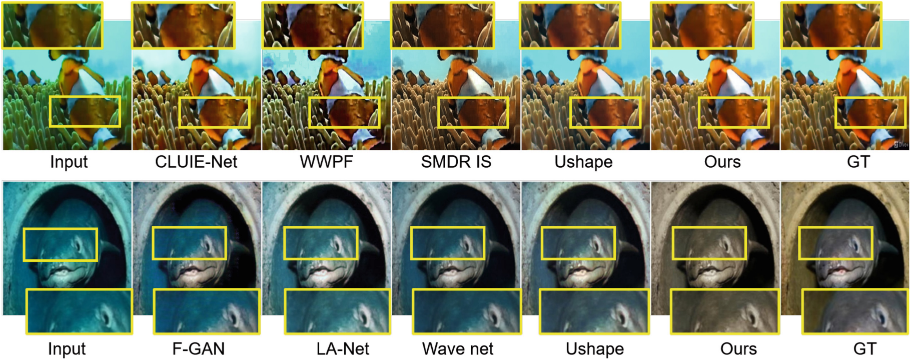


Fig. 5. Qualitative analysis of the proposed (Ours) and existing methods: LA-Net [18], F-GAN [10] CLUIE [16], Wave net [22], WWPF [31], SMDR IS [30], Ushape [21] on UIEB [15] (*top row*) and EUVP [10] (*bottom row*) database for underwater image enhancement. (Color figure online)

5.2 Quantitative Analysis

We have used the existing, LANET [18], Wave Net [22], CLUIE [16], WWPF [31], SMDR-IS [30], and Ushape [21] methods for analysis. Peak signal-to-noise ratio (PSNR), structural similarity index measure (SSIM), and learned perceptual image patch similarity (LPIPS) metrics are used for reference based analysis and is provided in Table 1. Along with this, we also have provided non-reference parameter based analysis in terms of average Underwater Image Quality Measurement (UIQM), Underwater Color Image Quality Evaluation (UCIQE), Underwater Image Colourfulness Measure (UICM), and Natural Image Quality Evaluator (NIQE) on the Color-checker and Sea-thru real-world datasets in Table 2 (*More analysis over UIEB, EUVP, UFO-120, and U-45 datasets is available in supplementary material*). The reference and non-reference results depict that the proposed method competes favorably for both synthetic and real world underwater image-enhancing tasks with recent state-of-the-art methods.

5.3 Qualitative Analysis

Our proposed method is qualitatively evaluated against previous state-of-the-art approaches on various synthetic datasets (EUVP and UIEB) and real-world datasets (Sea-thru and Color-checker) for enhancing underwater deteriorated images. The synthetic and Real-World underwater visual results on EUVP, UIEB are provided in Fig. 5 and 6. The Fig. 7 shows applicability of our proposed approach on higher-level computer vision application such as depth estimation applied on underwater degraded images and restored outputs with existing and proposed methods. From these provided visual results, it's clear that our proposed method produces real colors and clear images as compared to state-of-the-art methods. *More qualitative results on real-world and synthetic datasets are available in supplementary material.*

5.4 Ablation Studies

We analyse the contributions of each proposed module and UIEB dataset is used for ablation experimentation in terms of average PSNR and SSIM.

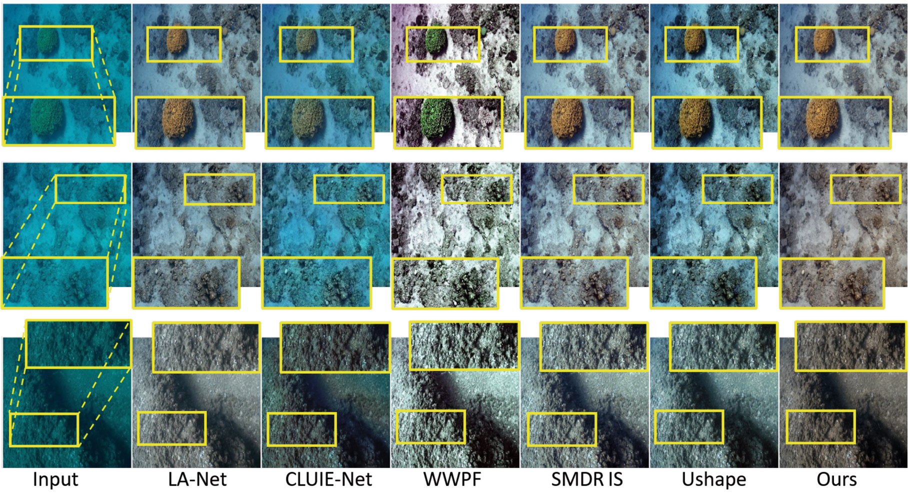


Fig. 6. Qualitative results comparison on sea-thru dataset with state-of-the-art (LA-Net [18], WaveNet [22], CLUIE [16], WWPF [31], SMDR-IS [30] and Ushape [21]) methods for underwater image enhancement. (Color figure online)

Multi-scale Feature Fusion Feed-Forward (MSF) Module: In the proposed MSF, we have utilized multi-scale convolution in the multi-scale feature

Table 3. Performance analysis on proposed MSF module (*Note: w/i:with and w/o: without, AFP: Adaptive frequency preserving*).

Setting	PSNR	SSIM
MSF w/o Multi-scale and w/o AFP	23.94	0.930
MSF w/i Multi-scale and w/o AFP	24.38	0.940
MSF w/o Multi-scale and w/i AFP	24.60	0.943
MSF w/i Multi-scale w/i AFP (Proposed)	**25.79**	**0.955**

Table 4. Performance analysis with various types of attention modules (*Note: SA: Self Attention*).

Attention Type	PSNR	SSIM
MDTA [29]	23.40	0.930
Depthwise Convolution SA	23.81	0.933
Deformable Convolution SA	24.42	0.937
FMDA (Proposed)	**25.79**	**0.955**

fusion feed forward module, after the adaptive frequency preserving (AFP) block to capture the fine details and improve contextual information. *Is this utilization of multi-scale convolution along with encompassing AFP block in MSF module effective to capture local and global contextual information.* The quantitative analysis with MSF w/o multi-scale and w/o AFP (normal feed-forward network), MSF w/i multi-scale w/o and AFP, MSF w/o multi-scale w/i AFP and MSF w/i multi-scale w/i AFP (proposed MSF) is given in the Table 3. It is evident from these results that the proposed MSF module is more effective as compared to other combinations of feed forward network.

Table 5. Performance analysis with various offset settings.

Offset Type	PSNR	SSIM
Modulated Deformable Offset	23.65	0.930
Spatially Attentive Deformable Offset	24.10	0.941
SCMDC Offset (Proposed)	**25.79**	**0.955**

Frequency Modulated Deformable Attention Module: Along with effectively capturing local contextual information, having long range dependencies with ability to accommodate according to variation in image is crucial task. To do this, we have proposed FMDA module. *Whether the proposed FMDA helps the network to capture contextual information effectively?* To examine this, we have performed the experimentation with MDTA [29], Depthwise convolution

Table 6. Analyzing the performance of various skip connection types (*Note: w/i:with and w/o: without*).

Skip Connection (SC) Type	PSNR	SSIM
Regular SC	24.26	0.931
SC w/i up-sample w/o AFP block	24.75	0.931
SC w/o up-sample w/i AFP block	24.76	0.934
SC w/i AFP and up-sample (Proposed)	**25.79**	**0.955**

SA, Deformable convolution SA and proposed FMDA. Quantitative analysis for above experimentation is provided in Table 4. Based on these findings, it is clear that the proposed FMDA is efficient for the UIE Task.

Spatio-channel Attentive Offset Extractor: In modulated deformable convolution, significant worry arises over offsets that may exceed their contextually relevant regions, resulting in the emergence of irrelevant features. To tackle this issue, we have proposed spatio-channel attentive offset extractor. *To scrutinize the efficiency of the proposed offset extractor over other offset extractor, we trained the proposed network with various offset extractor.* Here, the accuracy is analysed with modulated deformable offset, spatially attentive deformable and proposed SCMDC offset. A quantitative analysis can be found in Table 5 and (*see* Fig. 4), demonstrating the effectiveness of our proposed offset extractor. to focus on contextually relevant regions and color variations.

Adaptive Edge Preserving Skip Connection: Applying direct skip connection may led to traversing of degraded intermediate features from encoder to respective decoder. Thus, we have proposed AEP module to traverse the refine edge information from encoder to respective decoder. *Whether the proposed AEP module able to refine edge information for the underwater enhancement?* To analyze this, the accuracy of the proposed network is analysed with various combination of skip connections. The experimental quantitative results is shown in Table 6. From the Fig. 8 results, it's clear that our proposed AEP module in skip connection is more effective as compared to other modules.

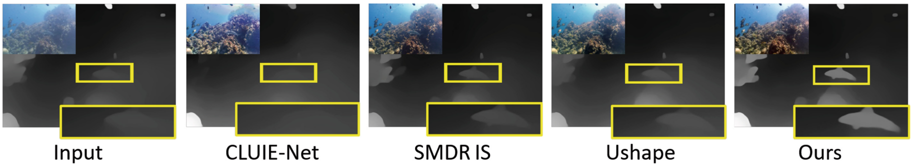


Fig. 7. Depth estimation analysis on input and enhanced images with existing CLUIE [16], SMDR-IS [30], Ushape [21] and proposed method (Ours). (Color figure online)

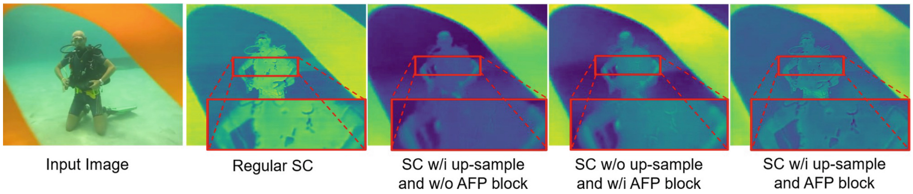

Input Image
Regular SC
SC w/i up-sample and w/o AFP block
SC w/o up-sample and w/i AFP block
SC w/i up-sample and AFP block

Fig. 8. Feature map visualization of various types of skip connections.

6 Conclusion

In this work, we have proposed novel frequency modulated deformable transformer for underwater image enhancement. In that, we proposed multi-scale feature fusion feed forward module for effective feature extraction. Further, the frequency modulated deformable transformer with spatio-channel attentive offset extractor is proposed for relevant contextual underwater image enhancement. Finally, we proposed adaptive edge-preserving module for propagating prominent edge features from the network's shallow layers to its deeper layers via skip connection. Experimental studies are carried out on both synthetic and real-world data sets using both reference and also non-reference parameters, and these are compared with recent state-of-the-art approaches. Numerous experimental analyses on synthetic and real-world images, along with a detailed ablation study scrutinize the effectiveness of the proposed method for underwater image enhancement.

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. Akkaynak, D., Treibitz, T.: Sea-thru: a method for removing water from underwater images. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 1682–1691 (2019)
2. Ancuti, C.O., Ancuti, C., De Vleeschouwer, C., Bekaert, P.: Color balance and fusion for underwater image enhancement. IEEE Trans. Image Process. **27**(1), 379–393 (2017)
3. Cho, S.J., Ji, S.W., Hong, J.P., Jung, S.W., Ko, S.J.: Rethinking coarse-to-fine approach in single image deblurring. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 4641–4650 (2021)
4. 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. Graph. Appl. **36**(2), 24–35 (2016)
5. 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)

6. Fan, Z., et al.: Mask attention networks: rethinking and strengthen transformer. arXiv preprint arXiv:2103.13597 (2021)
7. Fu, X., Liao, Y., Zeng, D., Huang, Y., Zhang, X.P., Ding, X.: A probabilistic method for image enhancement with simultaneous illumination and reflectance estimation. IEEE Trans. Image Process. **24**(12), 4965–4977 (2015)
8. Hitam, M.S., Awalludin, E.A., Yussof, W.N.J.H.W., Bachok, Z.: Mixture contrast limited adaptive histogram equalization for underwater image enhancement. In: 2013 International Conference on Computer Applications Technology (ICCAT), pp. 1–5. IEEE (2013)
9. 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: Schoeffmann, K., et al. (eds.) MMM 2018, Part I. LNCS, vol. 10704, pp. 453–465. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-73603-7_37
10. Islam, M.J., Xia, Y., Sattar, J.: Fast underwater image enhancement for improved visual perception. IEEE Robot. Autom. Lett. **5**(2), 3227–3234 (2020)
11. 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)
12. 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)
13. 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)
14. Li, C., Anwar, S., Porikli, F.: Underwater scene prior inspired deep underwater image and video enhancement. Pattern Recogn. **98**, 107038 (2020)
15. Li, C., et al.: An underwater image enhancement benchmark dataset and beyond. IEEE Trans. Image Process. **29**, 4376–4389 (2019)
16. Li, K., et al.: Beyond single reference for training: underwater image enhancement via comparative learning. IEEE Trans. Circuits Syst. Video Technol. (2022)
17. Liu, C., Shu, X., Pan, L., Shi, J., Han, B.: Multi-scale underwater image enhancement in RGB and HSV color spaces. IEEE Trans. Instrum. Meas. (2023)
18. Liu, S., Fan, H., Lin, S., Wang, Q., Ding, N., Tang, Y.: Adaptive learning attention network for underwater image enhancement. IEEE Robot. Autom. Lett. **7**(2), 5326–5333 (2022)
19. Liu, X., Lin, S., Chi, K., Tao, Z., Zhao, Y.: Boths: super lightweight network-enabled underwater image enhancement. IEEE Geosci. Remote Sens. Lett. **20**, 1–5 (2022)
20. Liu, Z., et al.: Swin transformer: hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 10012–10022 (2021)
21. Peng, L., Zhu, C., Bian, L.: U-shape transformer for underwater image enhancement. IEEE Trans. Image Process. **32**, 3066–3079 (2023)
22. 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)
23. Shen, Z., Xu, H., Luo, T., Song, Y., He, Z.: UdaFormer: underwater image enhancement based on dual attention transformer. Comput. Graph. **111**, 77–88 (2023)
24. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)

25. Tang, Y., Iwaguchi, T., Kawasaki, H., Sagawa, R., Furukawa, R.: Autoenhancer: transformer on U-net architecture search for underwater image enhancement. In: Proceedings of the Asian Conference on Computer Vision, pp. 1403–1420 (2022)
26. Wang, R., Zhang, Y., Zhang, J.: An efficient swin transformer-based method for underwater image enhancement. Multimedia Tools Appl. **82**(12), 18691–18708 (2023)
27. Wang, Y., Liu, H., Chau, L.P.: Single underwater image restoration using adaptive attenuation-curve prior. IEEE Trans. Circuits Syst. I Regul. Pap. **65**(3), 992–1002 (2017)
28. 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)
29. 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)
30. Zhang, D., Zhou, J., Guo, C., Zhang, W., Li, C.: Synergistic multiscale detail refinement via intrinsic supervision for underwater image enhancement. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 38, pp. 7033–7041 (2024)
31. Zhang, W., et al.: Underwater image enhancement via weighted wavelet visual perception fusion. IEEE Trans. Circuits Syst. Video Technol. (2023)
32. Zhu, X., Hu, H., Lin, S., Dai, J.: Deformable convnets v2: more deformable, better results. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 9308–9316 (2019)

Probing Attention-Driven Normalizing Flow Network for Low-Light Image Enhancement

Siddharth Singh[1], Nancy Mehta[2](✉), K. N. Prakash[3],
Santosh Kumar Vipparthi[1], and Subrahmanyam Murala[4]

[1] CVPR Lab, Indian Institute of Technology, Ropar, Rupnagar, India
[2] Vision Lab, CAIDAS & IFI, University of Wuerzburg, Würzburg, Germany
nancy.mehta@uni-wuerzburg.de
[3] SR Gudlavalleru Engineering College, Vijayawada, India
[4] CVPR Lab, School of Computer Science and Statistics, Trinity College Dublin, Dublin, Ireland

Abstract. Existing low-light image enhancement approaches based upon pixel-wise reconstruction losses are inadept at capturing the complex distribution of well-exposed images, resulting in residual noise, insufficient illuminance, and artifacts. Additionally, the mapping relationship between weakly-illuminated and normally exposed images is one-to-many, making low-light image enhancement a vastly ill-posed problem. In this work, we probe into this one-to-many relationship via an attention and frequency driven normalizing flow network by minimizing the negative log-likelihood loss. The proposed model comprises of two parts: a dual-attention-oriented frequency encoder network and an invertible network which inputs the conditional low-light images and changes the mapping of the complex distribution of well-light images to simpler Gaussian distribution. The proposed model not only utilizes the spatial information inherent in the image for improving the contrast but also extracts the frequency information for preserving the intricate details. To sum up, the distribution of the well-exposed images can be characterized better, and the overall enhancement mechanism becomes analogous to being restrained by a loss function which defines the manifold structure of natural images during the training. Detailed experiment analysis on a variety of challenging low-light images exemplifies the potency of the model and shows its primacy over the state-of-the-art in terms of enhanced quality and efficiency.

Keywords: Normalizing flow · Low-Light Image Enhancement · Frequency driven attention

1 Introduction

The visual quality of images is paramount in information transmission, significantly influencing human visual perception and the performance of computer vision systems, including autonomous driving [23], image segmentation [27], and

A. Antonacopoulos et al. (Eds.): ICPR 2024, LNCS 15332, pp. 137–151, 2025.
https://doi.org/10.1007/978-3-031-78125-4_10

object detection [35]. Nevertheless, due to the intrinsic properties of camera imaging mechanisms and the environmental conditions during photography, images captured in low-light conditions frequently suffer from low contrast, substantial noise, and poor color fidelity. Addressing these issues necessitates effective low-light image enhancement techniques, which are increasingly demanded in computer vision tasks. Contemporary deep learning-based enhancement methods typically employ pixel-wise loss functions in their network training to establish mappings between normally exposed and low-light images. However, this paradigm encounters notable challenges: firstly, pixel-to-pixel mappings are often constrained by the regression-to-the-mean issue, resulting in images that are undesired amalgamations of several targets, leading to under-exposed regions and artifacts. Secondly, the simplistic assumption underlying pixel-wise losses may fail to capture the visual distance between the enhanced image and the ground truth within the image manifold. While certain GAN-based approaches have mitigated this issue, they require meticulous tuning during training and are susceptible to overfitting the visual features of the training data.

Recent studies [16,21,33] have demonstrated the effectiveness of normalizing flow in learning conditional distributions rather than relying on basic pixel-wise loss, thereby overcoming the aforementioned limitations. Unlike traditional CNN-based methods [2,20], which learn deterministic mappings from low-light to well-exposed images, normalizing flow models map the image manifold to a latent distribution through a sequence of invertible and differentiable transformations. However, classical normalizing flows, which are biased towards learning graphical properties such as local pixel correlations [15] may fail to effectively model global image properties like color saturation, potentially undermining their performance in low-light image enhancement tasks. The proposed approach, by facilitating the construction of complex posterior distributions, overcomes this limitation by enhancing the modeling of structural details, illumination adjustment, and noise suppression, which are crucial for improving the quality of low-lit images.

We introduce an Attention-Driven Normalizing Flow network (ADNFNet), engineered to model the complex distributions of normally exposed images corresponding to low-light inputs. ADNFNet consists of two main components: a dual-attention-oriented frequency encoder for precise noise-free color map extraction and for the integration of global information into the latent space, and an invertible network for learning one-to-many mappings from low-light images to well-lit image distributions. In this framework, we refrain from using the standard Gaussian distribution as the prior for latent features, opting instead to use the illumination-invariant color map as the mean value of the prior distribution. The encoder is specifically designed to learn a one-to-one mapping to extract the color map, representing the intrinsic attributes of the scene that remain unchanged by illumination. Concurrently, the invertible network is crafted to learn a one-to-many mapping from low-light images to the distribution of normally exposed images. This design aims to achieve superior low-light image enhancement performance through the proposed framework. The principal contributions are as follows:

- An efficient attention-driven normalizing flow-based model that learns rich distributions through precise visual distance measurements, improving illumination adjustment and efficiently manage the noise/artifact suppression.
- A novel dual-attention-oriented frequency encoder module that minimizes color distortion and enhances saturation to extract an illumination-invariant color map.

Comprehensive experiments conducted on state-of-the-art enhancement methodologies substantiate the effectiveness of the proposed network. Through rigorous testing and comparison with existing techniques, we have demonstrated significant improvements in image quality, including enhanced contrast, reduced noise, and better color fidelity. Moreover, detailed ablation studies have been performed to isolate and validate the contribution of each individual module within the architecture. These studies confirm the rationality and necessity of each component, highlighting how they collectively contribute to the overall performance and robustness of the network.

2 Related Work

2.1 Low Light Enhancement

Early advancements in image quality enhancement primarily leveraged heuristic algorithms. For example, histogram equalization [10] effectively redistributes image brightness to enhance global contrast. Retinex theory-based methods [7,22] improve low-light images by decomposing them into reflectance and illumination components. The LIME algorithm [7] estimates the illumination intensity of each pixel and refines the initial illumination map using structural priors, thereby enhancing image quality. Despite their independence from training data, traditional methods often struggle with detail preservation and noise control. In recent years, deep learning-based methods have gained prominence due to their accuracy, robustness, and speed, setting new benchmarks in image enhancement tasks. Consequently, several methods with variations in architectural design were proposed. For example, where LLNet [20] deployed a deep autoencoder, multi-scale features were adopted in [25,26] for enlightening the image. The authors in [25] exemplifies the relationships between Retinex and convolutional neural networks (CNNs) via Gaussian kernels in a mutually reinforced manner. In addition, several approaches experimented with the overall training via different losses, *e.g.* L_1 loss [2], MSE [2,20], smoothness [29], and color loss [28]. The Restormer model [37] achieves high-resolution image restoration through a sophisticated Transformer architecture. Concurrently, Zhou *et al.* [42] address joint low-light enhancement and deblurring, introducing the extensive LOL-Blur dataset and demonstrating effectiveness on both synthetic and real-world data. Additionally, hybrid methods that integrate Retinex theory with deep learning [3,18,39] enhance images by optimizing reflectance and illumination components within the network. Unlike previous approaches that meticulously design different architectures or losses for end-to-end training, in this work, we intend to deploy attention-driven normalizing flow network for building the complex distribution that has shown to be proficient in generating images with better quality, lesser distortion, and artifacts.

2.2 Normalizing Flow

Normalizing flow transforms a simple probability distribution (*e.g.*, standard normal) into a complex distribution through a sequence of invertible and differentiable mappings [16]. This transformation allows exact computation of the probability density function (PDF) of a sample by reverting to the simple distribution. To ensure network invertibility and computational tractability, network layers must be meticulously designed to facilitate easy computation of the inversion and the Jacobian matrix determinant, which constrains the generative model's capacity. Consequently, various transformations have been developed to enhance the model's expressiveness, such as affine coupling layers [4], split and concatenation [4,5,14], permutation [14], and 1×1 convolution [14]. Conditional normalizing flows have been explored to bolster model expressiveness. Recently, conditional affine coupling layers [1,21,33] have been employed to strengthen the connection with conditional features, improving memory and computational efficiency. The development of normalizing flow has broadened its application scope significantly. For example, in [19], the authors generated faces with specific attributes, while Pumarola *et al.* [24] and Yang *et al.* [36] used conditional flow for point cloud generation. For the super-resolution tasks, the authors in [21], and [33] utilized conditional normalizing flow to generate high-resolution images from low-resolution inputs. Unlike other approaches for conditioning the probability distribution, the method incorporates both spatial and frequency domain features of the input.

3 Proposed Method

We propose an attention-driven normalizing flow (ADNFNet) framework to characterize the complex distribution of well-lit images. The overall paradigm of ADNFNet is demonstrated in Fig. 1, which embodies two key components: a dual attention oriented frequency (DAoFE) encoder and a series of invertible networks. DAoFE takes a low-light image (y_l) as input and outputs an illumination-invariant color map $g(y_l)$, and an invertible network maps a normally exposed image to a latent code z. The DAoFE component further consists of several dual-channel spatial attention modules (DCSAMs) which embodies residual channel attention blocks (RCABs), convolutional block attention modules (CBAMs), and spatial-frequency information refinement (SFIR) modules. In what follows, we illustrate in detail all the components.

3.1 Dual Attention Oriented Frequency Encoder

For boosting the generation of high-quality light invariant color maps, the concatenated feature maps of the low light image (y_l), and its equivalent histogram equalized image $h(y_l)$, and color map $C(y_l)$ are given as an input to

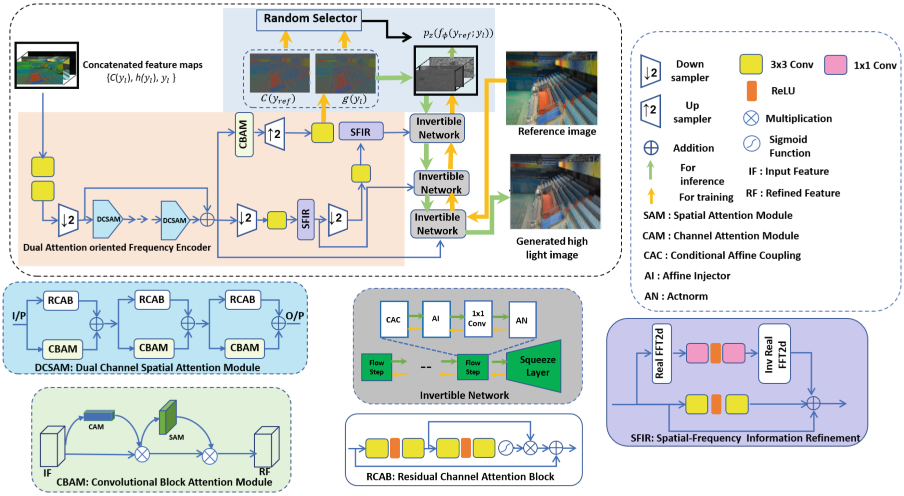


Fig. 1. Schematic illustration of the proposed ADNFNet. The proposed model comprises of a dual attention oriented frequency (DAoFE) encoder (light orange color) to extract the color map of the low light images and a series of invertible network for learning the distribution of well-lit images that are conditioned on a low-light image. A random selector is deployed for getting the mean value of latent variable z that follows the Gaussian distribution from the color map of the reference image, $C(y_{ref})$ or the extracted color map $g(y_l)$ from the low-light image via DAoFE. For training the exact likelihood of a high-light image (y_h) is maximized and for inference, we randomly select z from N$(g(y_l)$, 1) to generate multiple normally exposed images. (Color figure online)

the ADNFNet network. Basically, histogram equalization is used to enhance the global contrast of low-light images, making the histogram-equalized image more illumination invariant. By incorporating this image into the network's input, the network can more effectively handle areas that are excessively dark or bright. Further, inspired by Retinex theory [32], the color map is calculated as $C(y) = y/meanc(y)$, where meanc(.) computes the mean value of each pixel across RGB channels. This color map serves as a reflectance-like representation, with $C(y_l)$ and $C(y_{ref})$ maintaining consistency across different lighting conditions despite noise in $C(y_l)$.

For boosting the generation of high-quality light invariant color maps, the concatenated feature maps of the low light image (y_l), and its equivalent histogram equalized image $h(y_l)$, and color map $C(y_l)$ are first pre-processed via two 3×3 convolution layers for extracting the appropriate features as depicted in Fig. 1. The extracted features are thereafter given as input to the cascaded dual channel spatial attention module (DCSAM) which encompasses parallel residual channel attention blocks (RCAB) [41] and convolutional block attention module (CBAM) [34] as shown in Fig. 1 to fine-tune the color maps. This parallel design facilitates the extraction of the contextual associations by taking merit of the internal prior information among the concatenated feature maps and thus helps to subsequently enhance the textural details of the incoming low-light image. For making a trade-off between the local details for fine texture and global information for the overall brightness level, we deployed RCAB. CBAM is inserted to perform feature attention in both channel-wise and spatial-wise views. The essence of the dual attention in CBAM is to redistribute weights and to gather more global structural information inherited in the extracted feature maps.

To further enhance the global contextual learning capability of the proposed ADNFNet, we introduce the spatial-frequency information refinement (SFIR) module at two strategic locations, as depicted in Fig. 1. The design of the SFIR module incorporates both a spatial residual stream and a parallel channel-wise Fast Fourier Transform (FFT) stream. This module provides several key advantages. Firstly, by integrating the spatial and frequency domains, it leverages the benefits of both pixel-level and kernel-level features. This dual-domain representation fusion enables the network to capture more comprehensive and nuanced information from the input images. Specifically, the spatial stream processes the image details at the pixel level, while the FFT stream analyzes the image in the frequency domain, which is particularly effective for identifying repetitive patterns and textures. Secondly, the SFIR module effectively addresses the issue of blur and low semantic contrast in low-illuminated images. Low-light conditions often lead to poor contrast and noise, which can obscure important details. The frequency domain analysis in the SFIR module helps to filter out this noise, thereby enhancing the clarity and quality of the features extracted from the images. This results in a richer and cleaner set of features being passed on to the subsequent stages of the network.

Additionally, the SFIR module enriches the input features for the next invertible layer stage by removing noise and preserving essential details. The combination of spatial and frequency information ensures that the extracted features are robust and informative, facilitating better performance in the invertible network layers. In summary, the DAoFE component in the proposed ADNFNet framework learns a one-to-one mapping for extracting the color map $g(y_l)$ in both spatial and frequency domains. This approach yields a superior representation that is well-suited for conditioning in all flow layers of the invertible network, ultimately enhancing the network's ability to handle low-light image enhancement tasks.

3.2 Invertible Network

The invertible network in the framework is designed to handle one-to-many relationships due to the diverse illumination maps extracted from DAoFE network for the same scene. The overall objective is to capture the full conditional probability distribution $p_{y_{\text{ref}}|y_l}(y_{\text{ref}}|y_l, \phi)$ of well-lit images y_{ref} corresponding to low-light images y_l. The normalizing flow aims to parameterize this distribution via an invertible neural network f_ϕ.

The invertible network maps the well-exposed image y_{ref} into a latent variable $z = f_\phi(y_{\text{ref}}; y_l)$. This mapping must be invertible with respect to y_{ref} for any given y_l. Thus, the well-lit image y_{ref} can be reconstructed from the latent encoding

z as $y_{\text{ref}} = f_{\phi}^{-1}(z; y_l)$. By defining a distribution $p_z(z)$ in the latent space, the conditional distribution $p_{y_{\text{ref}}|y_l}(y_{\text{ref}}|y_l, \phi)$ can be implicitly defined by mapping samples $z \sim p_z$. The probability densities are explicitly computed using the change of variable theorem:

$$p_{y_{\text{ref}}|y_l}(y_{\text{ref}}|y_l, \phi) = p_z(f_\phi(y_{\text{ref}}; y_l)) \left|\det \frac{\partial f_\phi}{\partial y_{\text{ref}}}(y_{\text{ref}}; y_l)\right| \tag{1}$$

To achieve a tractable expression for this term, the invertible network f_ϕ is decomposed into sequential invertible layers. The network consists of three levels, each containing a squeeze layer and twelve flow steps, with each flow step comprising four distinct invertible layers. The careful design of these flow layers ensures a well-conditioned and tractable Jacobian determinant, minimizing the negative log-likelihood function. The key components of the network are:

- **Squeeze Layer:** Captures the incoming features from DAoFE at different scales by reshaping the feature map from $C \times H \times W$ to $4C \times \frac{H}{2} \times \frac{W}{2}$, increasing the network's receptive field.
- **Invertible 1×1 Convolution:** Functions similarly to a vanilla convolution layer with a kernel size of 1, allowing efficient determinant calculation.
- **Conditional Affine Coupling Layer:** Introduces the conditional feature $g(x_l)$ into the network, establishing a connection between low-light and normally exposed images. The operation is defined as:

  $$h_{i+1}^A = h_i^A; \quad h_{i+1}^B = \exp(\theta_i^s([h_i^A; z_i])) \cdot h_i^B + \theta_i^b([h_i^A; z_i]) \tag{2}$$

  where z_i is the conditional feature, and θ_i^s and θ_i^b are networks predicting scale and bias.
- **Affine Injector:** Strengthens the connection between the conditional feature and the well-lit image y_{ref}, defined as:

  $$h_{i+1} = \exp(\theta_i^s(z_i)) \cdot h_i + \theta_i^b(z_i) \tag{3}$$

- **Actnorm:** Performs channel-wise normalization via learned scaling and bias, similar to batch normalization.

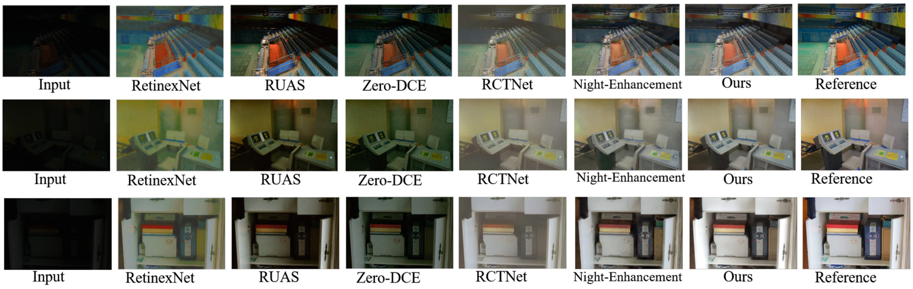


Fig. 2. Qualitative comparison with SoTA low-light enhancement approaches on LOL Dataset. The proposed ADNFNet effectively generate images that are visually closer to the reference, artifact-free, and more natural.

The networks θ_i^s and θ_i^b in the conditional affine coupling layer and affine injector layer consist of two shared convolutional layers with 64 hidden channels and PReLU activation, followed by a convolutional layer to predict the scale and bias.

4 Experiment

4.1 Experimental Settings

Following the settings in [30], the size of all patches has been set at 160×160, and the overall batch size is set at 16. The model has been trained for $3{\times}10^4$ iterations and the overall learning rate is decremented by a factor of 0.5 at 1.5×10^4, 2.25×10^4, $2.7{\times}10^4$, and 2.85×10^4 iterations for LOL dataset. For VE-LOL dataset, the model is trained for 4×10^4 iterations, and the overall learning rate is reduced by a factor of 0.5 at $2{\times}10^4$, 3×10^4, 3.6×10^4, and 3.8×10^4 iterations. In the entire training, we utilized Adam as the optimizer with a learning rate of $5{\times}10^{-4}$. In order to efficiently mark the properties of well-lit images, we deploy the maximum likelihood estimation for estimating the learnable parameter, ϕ in Eq. 1 and we aim at reducing the negative log-likelihood (NLL) loss for training samples (y_l, y_{ref}).

$$L(\phi; y_l, y_{ref}) = -\log p_{y_{ref}|y_l}(y_{ref}|y_l, \phi); \tag{4}$$

Using Eq. 1, the above equation reduces to minimizing the following loss:

$$= -\log p_z(f_\phi(y_{ref}; y_l)) - \log|\det \tfrac{\partial f_\phi}{\partial y}(y_{ref}; y_l)| \tag{5}$$

4.2 Experiments on LOL

We evaluate the proposed ADNFNet using 15 images for testing and 485 images for training from LOL dataset [32]. For quantitative comparison, three metrics

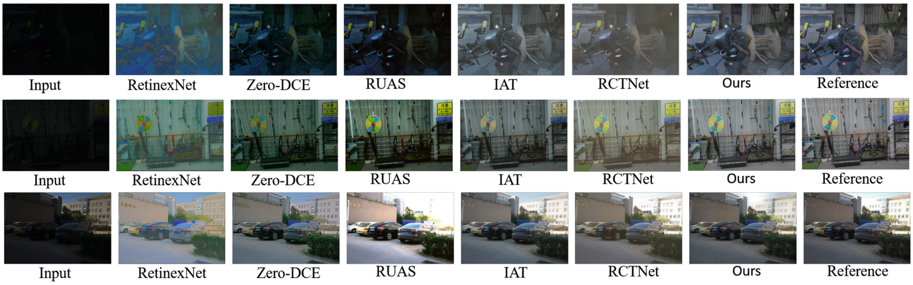


Fig. 3. Qualitative comparison with SoTA low-light enhancement approaches on LOL-v2 Dataset. THe proposed ADNFNet effectively synthesizes images that exhibit enhanced visual fidelity to the reference and possess a more natural appearance in cross dataset evaluation settings.

are used, PSNR, SSIM [31], and LPIPS [38]. As indicated in Table 1, the proposed approach considerably surpasses all other competitors. Higher PSNR values indicate the method's superior capability in suppressing artifacts and accurately recovering color information. Improved SSIM values demonstrate that the approach excels in preserving structural information with high-frequency details. Furthermore, the method achieves the best performance in terms of LPIPS, a metric designed for human perception, indicating a closer alignment with human visual quality. The visual results in Fig. 2 further demonstrate that ADNFNet exposes more image details, and achieves more natural, and artifact-free results.

4.3 Experiments on VE-LOL

We further investigate the proposed method on VE-LOL dataset [17] in order to more accurately assess its effectiveness and generalizability. It is a sizeable dataset with 2500 paired photos having diverse themes and scenes, making it beneficial for evaluation in a cross-dataset manner.

Table 1. Quantitative comparison in terms of PSNR, SSIM and LPIPS on the LOL Dataset. ↑ (↓) denotes that, larger (smaller) values generate finer quality.

Methods	PSNR ↑	SSIM ↑	LPIPS ↓
RetinexNet [32]	16.77	0.56	0.47
Zero-DCE [6]	14.86	0.54	0.33
EnlightenGAN [11]	17.48	0.65	0.32
RUAS [18]	18.23	0.72	0.35
RCTNet [13]	22.67	0.79	0.23
Night-Enhancement [12]	21.52	0.76	0.25
Retinexformer [3]	25.16	0.84	0.18
GSAD [9]	**25.75**	0.82	0.16
Ours	25.70	**0.91**	**0.14**

1. **Cross-dataset evaluation**: We initially assess the universality of the proposed approach in a cross-dataset manner, *i.e.* training via LOL dataset and testing on VE-LOL dataset. The quantitative and qualitative results in Table 2 and Fig. 3 signify that the approach clearly beats the competing SoTA methods on all the metrics and outcome images with less noise and improved color saturation. The generated images via ADNFNet preserve more intricate details as compared to the other approaches.
2. **Intra-dataset evaluation**: For additional evaluation of the performance of ADNFNet, we analyze it with other SoTA methods in an intra-dataset manner, where we retrain on VE-LOL dataset and its corresponding test set is

Table 2. Quantitative comparison on VE-LOL dataset where all the models are trained on the training dataset of LOL.

Methods	PSNR ↑	SSIM ↑	LPIPS ↓
RetinexNet [32]	14.68	0.53	0.64
DeepUPE [28]	13.19	0.49	0.46
KinD [40]	18.42	0.77	0.29
Zero-DCE [6]	21.12	0.77	0.25
KinD++ [39]	17.63	0.79	0.23
EnlightenGAN [11]	20.43	0.79	0.24
Ours	**24.47**	**0.87**	**0.17**

Table 3. Quantitative comparison on the VE-LOL dataset where the models are trained on the training set of VE-LOL.

Methods	PSNR ↑	SSIM ↑	LPIPS ↓
KinD [40]	22.15	0.85	0.26
Zero-DCE [6]	20.54	0.78	0.33
Ours	**26.37**	**0.92**	**0.13**

used for reporting performance. From Table 3 we can observe that the approach exhibits comparable performance. Meanwhile, it can be seen that with more diverse data, all the compared metrics have improved in comparison to the model trained on LOL dataset.

4.4 Ablation Study

Here is the detailed ablation study where all the evaluations have been meticulously carried out on the LOL dataset to ensure consistency and reliability of the results.

Table 4. Ablation study on different components of ADNFNet.

Modules	A1	A2	A3	A4	A5
Baseline	✓	✓	✓	✓	✓
DCSAM		✓	✓	✓	✓
SFIR			✓		✓
CBAM				✓	✓
PSNR	**22.12**	**25.09**	**25.27**	**25.35**	**25.70**

Effectiveness of the Proposed Components: In this section, we demonstrate the importance of each module in the proposed ADNFNet as shown in Tables 4 and5. For the baseline (A1), we deployed simple residual block [8] in place of DCSAM, channel attention [41] in place of CBAM, and no SFIR module. After appending each proposed component sequentially into the baseline as depicted in Table 4, there seems to be a consistent improvement in the overall performance on the LOL dataset. It clearly indicates that each module is proficient in performing the dedicated task. Overall, the proposed ADNFNet (A5) attains a captivating performance gain of 3.58 dB over the baseline (A1). After confirming the validity of the proposed components, we also analyze the effect of the different configurations of the two main components in the proposed DAoFE. From Table 5, it is clear that the parallel combination of the proposed components exhibits a performance gain of 5.63 and 1.44 dB in comparison to deploying only RCAB, and CBAM blocks, respectively.

Table 5. Ablation study on two main components of DCSAM.

Components	PSNR ↑	SSIM ↑
RCAB	20.07	0.74
CBAM	24.26	0.90
RCAB + CBAM (series)	25.57	0.91
RCAB + CBAM (parallel)	**25.70**	**0.91**

Effectiveness of Different Latent Distributions: The latent feature z follows the probability density function (PDF):

$$f_z(z) = \frac{1}{\sqrt{2\pi}} \exp\left(- \frac{(x - r(C(y_{\mathrm{ref}}), g(y_l)))^2}{2} \right)$$

where $C(y_{\mathrm{ref}})$, and $g(y_l)$ denotes the color map of the well-lit image, and $g(y_l)$ denotes the illumination invariant color map extracted from the encoder (DAoFE) and $r(a, b)$ is a random selection function defined as:

$$r(a, b) = \begin{cases} a & \beta \leq p \\ b & \beta > p \end{cases} \quad \text{with} \quad \beta \sim U(0,\ 1)$$

The hyper-parameter p is set to 0.2 for all experiments. To evaluate the effectiveness of the proposed illumination-invariant color map and various hyper-parameters p, we tested them using the LOL dataset (Wei et al. 2018). The results, shown in Table 6, indicate that the proposed model with the newly designed color map achieves better PSNR values. Additionally, the higher SSIM and LPIPS values demonstrate that the color map improves color and brightness consistency.

Table 6. The overall effect of distinct latent distributions on LOL

Latent Space Distribution	PSNR ↑	SSIM ↑	LPIPS ↓
ADNFNet (w/o color map)	25.11	0.91	0.15
ADNFNet (w color map, $p = 0.5$)	24.91	0.91	0.16
ADNFNet (w color map, $p = 0.2$)	**25.70**	**0.91**	**0.14**

5 Conclusion

In this paper, we introduce a sophisticated framework for low-light image enhancement leveraging a novel normalizing flow model. Unlike conventional techniques that rely on pixel-wise reconstruction losses and deterministic processes, the proposed approach utilizes negative log-likelihood (NLL) loss with low-light images/features as conditions. This inherently allows for superior characterization of structural context and a more accurate measurement of visual distance within the image manifold. Furthermore, the method exploits attention mechanisms to effectively capture contextual relationships and frequency information, which enhances the modeling of complex conditional distributions of normally exposed images. Consequently, this leads to superior low-light enhancement, characterized by well-exposed illumination, reduced noise and artifacts, and enriched color fidelity. Experimental evaluations on established benchmark datasets demonstrate that the framework achieves superior quantitative and qualitative performance compared to state-of-the-art methodologies, thus validating its efficacy and robustness.

Acknowledgments. This research was supported by the DST-SERB, India, under Grant CRG/2022/006876. The authors would also like to express their gratitude to the members of the CVPR Lab for their invaluable support in this work.

References

1. Ardizzone, L., Lüth, C., Kruse, J., Rother, C., Köthe, U.: Guided image generation with conditional invertible neural networks. arXiv preprint arXiv:1907.02392 (2019)
2. Cai, J., Gu, S., Zhang, L.: Learning a deep single image contrast enhancer from multi-exposure images. IEEE Trans. Image Process. **27**(4), 2049–2062 (2018)
3. 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)
4. Dinh, L., Krueger, D., Bengio, Y.: Nice: non-linear independent components estimation. arXiv preprint arXiv:1410.8516 (2014)
5. Dinh, L., Sohl-Dickstein, J., Bengio, S.: Density estimation using real NVP. arXiv preprint arXiv:1605.08803 (2016)

6. Guo, C., et al.: 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)
7. Guo, X., Li, Y., Ling, H.: Lime: low-light image enhancement via illumination map estimation. IEEE Trans. Image Process. **26**(2), 982–993 (2016)
8. 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)
9. Hou, J., Zhu, Z., Hou, J., Liu, H., Zeng, H., Yuan, H.: Global structure-aware diffusion process for low-light image enhancement. In: Advances in Neural Information Processing Systems, vol. 36 (2024)
10. Ibrahim, H., Kong, N.S.P.: Brightness preserving dynamic histogram equalization for image contrast enhancement. IEEE Trans. Consum. Electron. **53**(4), 1752–1758 (2007)
11. Jiang, Y., et al.: EnlightenGAN: deep light enhancement without paired supervision. IEEE Trans. Image Process. **30**, 2340–2349 (2021)
12. Jin, Y., Yang, W., Tan, R.T.: Unsupervised night image enhancement: when layer decomposition meets light-effects suppression. In: Avidan, S., Brostow, G., Cissé, M., Farinella, G.M., Hassner, T. (eds.) ECCV 2022, Part XXXVII. LNCS, vol. 13697, pp. 404–421. Springer, Cham (2022). https://doi.org/10.1007/978-3-031-19836-6_23
13. Kim, H., Choi, S.M., Kim, C.S., Koh, Y.J.: Representative color transform for image enhancement. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 4459–4468 (2021)
14. Kingma, D.P., Dhariwal, P.: Glow: generative flow with invertible 1x1 convolutions. Advances in Neural Information Processing Systems, vol. 31 (2018)
15. Kirichenko, P., Izmailov, P., Wilson, A.G.: Why normalizing flows fail to detect out-of-distribution data. In: Advances in Neural Information Processing Systems, vol. 33, pp. 20578–20589 (2020)
16. Kobyzev, I., Prince, S.J., Brubaker, M.A.: Normalizing flows: an introduction and review of current methods. IEEE Trans. Pattern Anal. Mach. Intell. **43**(11), 3964–3979 (2020)
17. Liu, J., Xu, D., Yang, W., Fan, M., Huang, H.: Benchmarking low-light image enhancement and beyond. Int. J. Comput. Vision **129**, 1153–1184 (2021)
18. 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)
19. Liu, R., Liu, Y., Gong, X., Wang, X., Li, H.: Conditional adversarial generative flow for controllable image synthesis. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 7992–8001 (2019)
20. Lore, K.G., Akintayo, A., Sarkar, S.: LLnet: a deep autoencoder approach to natural low-light image enhancement. Pattern Recogn. **61**, 650–662 (2017)
21. Lugmayr, A., Danelljan, M., Van Gool, L., Timofte, R.: SRFlow: learning the super-resolution space with normalizing flow. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J.-M. (eds.) ECCV 2020, Part V. LNCS, vol. 12350, pp. 715–732. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-58558-7_42
22. Ng, M.K., Wang, W.: A total variation model for retinex. SIAM J. Imag. Sci. **4**(1), 345–365 (2011)

23. Pham, L.H., Tran, D.N.N., Jeon, J.W.: Low-light image enhancement for autonomous driving systems using driveretinex-net. In: 2020 IEEE International Conference on Consumer Electronics-Asia (ICCE-Asia), pp. 1–5. IEEE (2020)
24. Pumarola, A., Popov, S., Moreno-Noguer, F., Ferrari, V.: C-flow: conditional generative flow models for images and 3d point clouds. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 7949–7958 (2020)
25. Shen, L., Yue, Z., Feng, F., Chen, Q., Liu, S., Ma, J.: MSR-net: low-light image enhancement using deep convolutional network. arXiv preprint arXiv:1711.02488 (2017)
26. Tao, L., Zhu, C., Xiang, G., Li, Y., Jia, H., Xie, X.: LLCNN: a convolutional neural network for low-light image enhancement. In: 2017 IEEE Visual Communications and Image Processing (VCIP), pp. 1–4. IEEE (2017)
27. Wang, L.W., Liu, Z.S., Siu, W.C., Lun, D.P.: Lightening network for low-light image enhancement. IEEE Trans. Image Process. **29**, 7984–7996 (2020)
28. Wang, R., Zhang, Q., Fu, C.W., Shen, X., Zheng, W.S., Jia, J.: Underexposed photo enhancement using deep illumination estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 6849–6857 (2019)
29. Wang, Y., et al.: 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)
30. Wang, Y., Wan, R., Yang, W., Li, H., Chau, L.P., Kot, A.: Low-light image enhancement with normalizing flow. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 36, pp. 2604–2612 (2022)
31. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: from error visibility to structural similarity. IEEE Trans. Image Process. **13**(4), 600–612 (2004)
32. Wei, C., Wang, W., Yang, W., Liu, J.: Deep retinex decomposition for low-light enhancement. arXiv preprint arXiv:1808.04560 (2018)
33. Winkler, C., Worrall, D., Hoogeboom, E., Welling, M.: Learning likelihoods with conditional normalizing flows. arXiv preprint arXiv:1912.00042 (2019)
34. Woo, S., Park, J., Lee, J.-Y., Kweon, I.S.: CBAM: convolutional block attention module. In: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (eds.) ECCV 2018. LNCS, vol. 11211, pp. 3–19. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-01234-2_1
35. Xu, X., Wang, S., Wang, Z., Zhang, X., Hu, R.: Exploring image enhancement for salient object detection in low light images. ACM Trans. Multimedia Comput. Commun Appl. (TOMM) **17**(1s), 1–19 (2021)
36. Yang, G., Huang, X., Hao, Z., Liu, M.Y., Belongie, S., Hariharan, B.: Pointflow: 3D point cloud generation with continuous normalizing flows. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 4541–4550 (2019)
37. 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)
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. Zhang, Y., Guo, X., Ma, J., Liu, W., Zhang, J.: Beyond brightening low-light images. Int. J. Comput. Vision **129**, 1013–1037 (2021)

40. 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)
41. Zhang, Y., Li, K., Li, K., Wang, L., Zhong, B., Fu, Y.: Image super-resolution using very deep residual channel attention networks. In: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (eds.) ECCV 2018. LNCS, vol. 11211, pp. 294–310. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-01234-2_18
42. Zhou, S., Li, C., Change Loy, C.: Lednet: joint low-light enhancement and deblurring in the dark. In: Avidan, S., Brostow, G., Cissé, M., Farinella, G.M., Hassner, T. (eds.) ECCV 2022. LNCS, vol. 13666, pp. 573–589. Springer, Cham (2022). https://doi.org/10.1007/978-3-031-20068-7_33

A Novel Encoder-Decoder Network with Multi-domain Information Fusion for Video Deblurring

Peiqi Xie, Jinhong He, Chengyun Song, and Minglong Xue(✉)

Chongqing University of Technology,Chongqing 400054, China
{paige,hejh}@stu.cqut.edu.cn, xueml@cqut.edu.cn

Abstract. Due to various challenging conditions during video recording, such as camera shake and out-of-focus issues, video deblurring remains a difficult problem. To address this, we propose the Spatial-Temporal Frequency domain Fusion network (STFFNet) and improve the network from three key aspects. Firstly, we introduce the Encoder-Decoder idea to create a novel backbone to combine global and local features effectively. Secondly, a new feature fusion module that focuses on the differences between frames is proposed to help better deblur the current frame. Finally, STFFNet introduces a Frequency Domain Converter (FDC) to transform the image information from the spatial domain to the frequency domain, enhancing image restoration by narrowing the gap between the deblurring and ground truth images in the frequency domain. Experimental results demonstrate that the proposed method achieves state-of-the-art deblurring performance on benchmark datasets. The code is available at: https://github.com/Paige-Norton/STFFNet.

Keywords: Video deblurring · Multi-domain Fusion · Differential Amplifier

1 Introduction

Video recordings often encounter quality problems caused by various factors, such as out-of-focus and camera shake, which can seriously affect the performance of downstream tasks, such as detection and tracking. Enhancing the quality of model recovery thus becomes a pressing problem for video-deblurring tasks. In the deblurring task, a blurred image is usually seen as the result of a blurring kernel acting on a sharp image. Earlier, researchers needed to design features to estimate sharp images manually. It was essential to develop image deblurring research, but it performed poorly in more complex scenarios.

Deep learning technology has brought about new solutions, and some novel approaches [1–6] have been proposed. Using deep learning technology, researchers can save the extra overhead of manually designing features and give full authority to the network model. In the deblurring task of video, utilizing the front and

A. Antonacopoulos et al. (Eds.): ICPR 2024, LNCS 15332, pp. 152–166, 2025.
https://doi.org/10.1007/978-3-031-78125-4_11

back frame images becomes the key. The CNN-based methods [7,8] deal with the problem by stacking neighbouring frames as input to the model. This method often makes it challenging to learn the time-domain information of the input data, resulting in a limited modelling effect. The RNN-based methods [9–12] use a recurrent neural network architecture to process the input information frame by frame, and the spatio-temporal information between frames is better utilized. However, due to limited information exchange between frames in this approach, information extraction lacks feedback and interaction. At the same time, there is a significant difference in representation between the blurred and sharp images in the frequency domain, as shown in Fig. 1. Whereas video deblurring algorithms usually focus on utilizing spatial and time domain information, they lack the utilization of frequency domain information. Therefore, it is very important to study how to use frequency domain information.

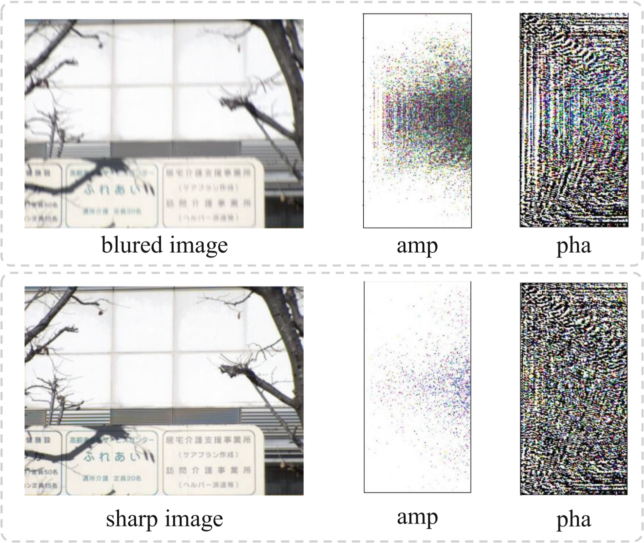


Fig. 1. The comparison chart. It shows the difference between blur images and ground-truth images in terms of amplitude(amp) and phase(pha) information.

This paper proposes an efficient video deblurring network that incorporates information from the frequency domain called the Spatio-Temporal Frequency Domain Fusion Network (STFFNet). The new backbone proposed for this network can decode each frame by combining all the features of the frame sequence so that the decoded features can contain global and local information. Also, we propose a new feature fusion module to selectively refer to frames with different degrees of blurring and pay more attention to the different parts of blurring. To utilize the difference between a blurred image and a sharp image in the frequency domain, we introduced a Frequency Domain Converter (FDC). It allows us to transform the spatial domain into the frequency domain and compute this difference in the frequency domain. In conclusion, we construct the network in this paper in three main ways to get more competitive deblurring performance:

- We propose a novel network structure as the backbone of STFFNet. This backbone can utilize the spatio-temporal information of video frames to provide a better basis for image recovery.
- We propose a novel data fusion module called Difference Amplification Blocks (DABs). This module focuses more on the blurring differences between frames and devotes more attention to the difference part.
- We introduce frequency domain information to minimize the content detail differences between the deblurring and ground-truth images.
- The experimental results show that our proposed method reaches the state-of-the-art on the benchmark datasets. Our method achieves the best visual and quantitative results.

2 Related Work

2.1 Video Deblurring

Unlike image-deblurring tasks [9,13,14], video-deblurring tasks correlate strongly with data. More than image work, video work needs to investigate how to utilize the spatio-temporal information among the data. In the early development of deep learning technology, researchers mainly invested in CNN-based methods [7,8,15]. Su et al. [7] proposed the first end-to-end data-driven video deblurring method, which aligns the video frame data by frame-by-frame single-response alignment or optical flow alignment and lets the pairs of subsequent frame data go through a model to estimate the center blurring frame. Wang et al. [8] use deformable convolution to achieve frame alignment at the feature level in a coarse to fine manner and propose a temporal and spatial attention (TSA) fusion module to emphasize essential features for subsequent recovery. As a result of these proposed modules, the EDVR achieves a huge advantage in the face of motion-heavy data. However, the attractive CNN-based approach stacks the input data, which does not use the time domain information well.

Thanks to the excellent performance shown by RNN in processing time-series signals, this has attracted the attention of another group of researchers [9,10,16,17]. Zhou et al. [10] proposed a spatio-temporal variant of RNN for video deblurring. The RNN is usually used to utilize the previous frames efficiently. Similarly, Park et al. [9] proposed to recycle previous feature information in each iteration and used an incremental temporal training procedure, not to train from the worst blurred image to the ground-truth image, but to gradually train from the image with higher blurring to the image with lower blurring. However, the structure of RNN still needs to be improved for information transfer and extraction, and its inability to effectively utilize global information restricts feature extraction. More advanced networks (e.g. Bi-LSTM [18], GRU [19], Seq2Seq [20]) have been proposed in natural language processing to extract more advanced features.

Inspired by these, the focus of this paper is on how to combine these advanced networks with video deblurring tasks.

2.2 The Frequency Domain in the Image

In the early days of image work, researchers usually focused on the spatial domain of the image for pixel-level processing. Continuing with the video work, the time domain information is introduced into the study because of the natural temporal connection between frames. Meanwhile, frequency domain analysis is an essential method in image work. The image can be better analyzed and understood by transforming the image from the time domain to the frequency domain. For example, image blurring results from insufficient high-frequency components in the image, and some blurring can be eliminated by increasing the high-frequency components or decreasing the low-frequency components in the frequency domain. Another example is that images are sometimes affected by recurring regular periodic noise, which has a specific frequency, so a frequency domain filtering approach can be taken to filter out the corresponding noise frequency, thus eliminating the periodic noise. Cai et al. [21] proposed that FDIT decompose an image into high-frequency and low-frequency features, using the former to capture object structures similar to identity recognition, thus achieving better image translation. Jiang et al. [22] used fast fourier convolution [23] to expand the network sensing field and enhance the network perception, enhancing the network generalization performance and reducing computational cost.

3 Methods

To recover the video frames to better results, STFFNet is proposed in this paper. It consists of three parts, and the framework is shown in Fig. 2. Firstly, every five successive blurred video frames extract the feature f for each frame using backbone (Fig. 2 (a)). Secondly, pass f to the feature fusion network (Fig. 2 (b)) to blend the five features f into mixed features F, and use the reconstructor (Fig. 2 (c)) to recover the mixed features F into a sharp image. Finally, the loss is passed back to guide the following process. In the following section, we describe the main components of our method in detail.

3.1 A Novel Backbone Network

In video work, the time domain information between frames is crucial. With the development of deep learning technology, RNN-based methods have replaced CNN-based methods as mainstream methods. However, this paper argues that global features can be incorporated more obviously in the RNN propagation process to improve the global nature of the generated features. Inspired by Seq2Seq and Bi-LSTM, we propose a novel backbone network named Decoding Every Frame (DEF). It uses the Encoder-Decoder idea, and it is shown in Fig. 2(a). The architecture is designed as follows: sequential encoding of video frames to

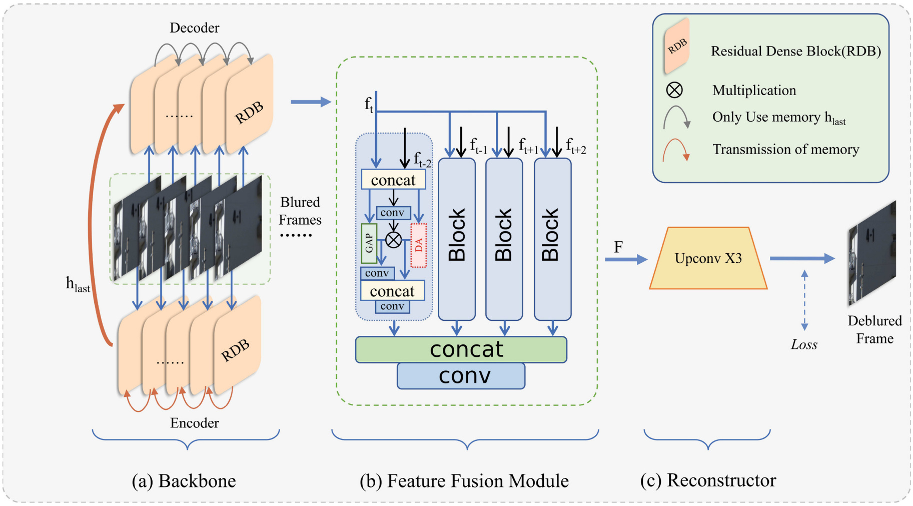


Fig. 2. The overall architecture of STFFNet. It contains three main components:(a). A novel backbone network called DEF to extract features of video sequences. (b). Our proposed feature fusion module is named DABs. It focuses on processing the differences between frames. (c). A reconstructor is designed to transform features into images.

obtain global features and decoding of each video frame only by the global features to obtain local features that contain global information. More specifically, sequential frames are first sequentially encoded in memory information h by the RDB cell and passed to the following frames. Then, the last memory information, h, is sequentially decoded by the RDB cell in combination with the video frames to obtain the feature f. With this framework, more advanced features are extracted.

$$h_i = \begin{cases} \emptyset & i = 0 \\ Encoder(x_i, h_{i-1}) & i \in [1,5] \end{cases}, \tag{1}$$

$$f_i = Decoder(x_i, h_{last}),\ i \in [1,5]. \tag{2}$$

In this architecture, the memory information h_i is generated in the encoder stage (where i stands for the ith frame and a video sequence is set to 5 frames). When in the first frame of the video, $i = 0$, and h_i is empty, then h_i is Encoder with each frame of the image. h_i generated in the last frame, denoted by h_{last}, contains the global information of the sequence of frames, formulized as Eq. 1. The global information h_{last} generated by the Encoder is utilized in the Decoder stage to decode the feature f_i with the ith frame respectively, and no more h_i is generated, formulized as Eq. 2. Each decoding process is independent and does not communicate with each other. Compared to the Seq2Seq network, it removes the h_i that is communicated between frames in the decoder phase while the encoder phase remains unchanged, and the Decoder retains only the output feature f_i. This change is experimentally driven, and this paper validates the effectiveness of this treatment in an ablation study. This paper argues that this is because the present frame already contains enough information to transfer. However, Seq2Seq passes too much redundant information in this task, further degrading network performance and increasing the difficulty of fitting.

3.2 Feature Fusion Module

After obtaining the features f_i of each frame, An urgent problem is the effective use of these features. In this paper, video deblurring is viewed as a problem of restoring a sharp frame from a sequence of blurred frames. Different frames have different effects on the reduction effect and cope with different attention during processing. In this paper, inspired by the GSA [24] module, we propose a new data fusion method called Difference Amplification Blocks (DABs), which consists of a fusion of multiple blocks, and the structure is shown in Fig. 2(b). The five features generated by DEF are fused here. The center frame features and the other four frame features are combined into four groups of feature sets, respectively, which go to Block to learn the differences between each set of features. Then, the four groups of feature differences are concatenated and convolved to the appropriate dimension to obtain the hybrid feature F, which is passed to the next step.

In contrast to the GSA structure, the DABs split a branch parallel to the GAP in each block, which uses the Differential Amplifier (DA) proposed in this paper to centralize the blurring differences between frames. The DA structure is shown in Fig. 3.

Specifically, the difference between a pair of frames DA(f_j)(where j denotes the jth pair of frames and $j \in [1, 4]$) is expressed as a pair of characteristic features of noncenter frames f_j minus characteristic of the center frame to obtain the difference information (expressed as differ(f_j)) and the product of the convolution of f_j, formulized as Eq. 4. DA(f_j) and the result after convolution of the GAP branch constitutes a block, four pairs of feature group f_j after the DA block will form the total features F, formulized as Eq. 5.

$$f_j = Concat(f_{t\pm k}, f_t), \tag{3}$$

$$DA(f_j) = Conv(f_j) \times differ(f_j), \tag{4}$$

$$F = \sum_{j=1}^{4} Concat + Conv\left[GAP(f_j) \times Conv(f_j), DA(f_j)\right], \tag{5}$$

where t denotes the center frame of the frame sequence and $t \in [3, N^+ - 2]$, where k denotes the number of frames that differ from the center frame and $k \in [1, 2]$.

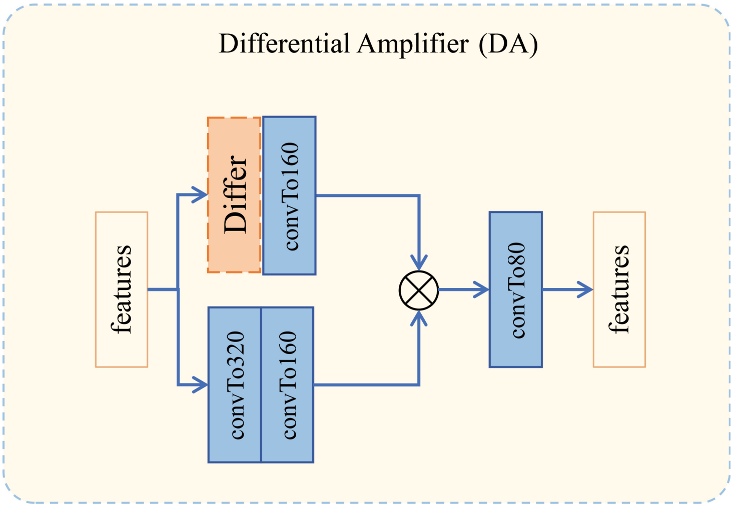


Fig. 3. The structure of a Differential Amplifier (DA). It is designed to focus on the differences between video frames. The structure's core is the Differ section, which focuses on the difference between the center video frame and the other frames.

3.3 Frequency Domain Guidance

Blurring or not blurring is expressed as a significant difference in the frequency domain, as shown in Fig. 1. It is usually considered that the frequency domain information can be utilized to guide image restoration from an additional perspective. In image processing, the amplitude usually indicates the magnitude of the contribution of each spatial frequency in the image. High amplitude usually corresponds to high-frequency portions of the image, such as edges or textures, which often contain detailed information in the picture. Low amplitude, on the other hand, corresponds to smooth regions or low-frequency portions. Blurred images tend to have weaker high-frequency information compared to sharp images. Then, increasing the high-frequency information of the recovered image while reducing the influence of low-frequency blurred information is another focus of this paper. This paper uses the Frequency Domain Converter (FDC) to introduce information in the frequency domain, the framework shown in Fig. 4. The predicted images and ground-truth are transformed into amplitude(amp) and phase(pha) information by FDC, respectively. The final disparity is obtained by calculating the Manhattan distance between the predicted images and ground-truth in amplitude and phase.

Formulized as Eq. 6, specifically using the extraction of amplitude and phase information for the predicted image and ground-truth, expressed as $pre_{amp}, pre_{pha}, gt_{amp}, gt_{pha}$, respectively, and then calculating the manhattan of the difference between the predicted image and the ground-truth in amplitude

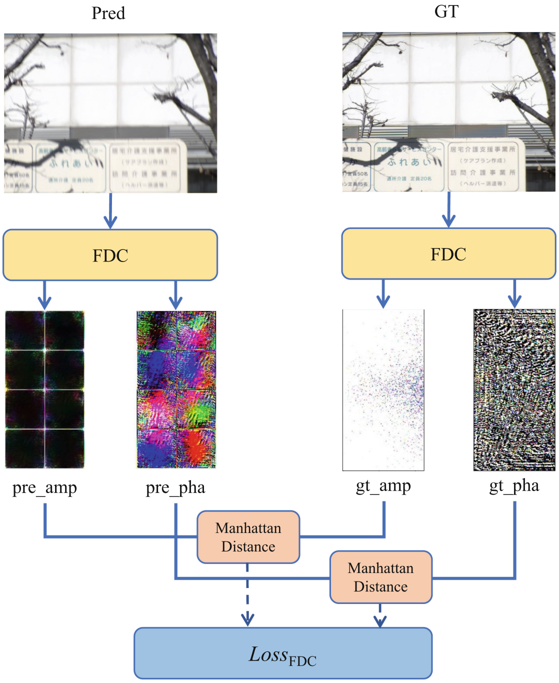


Fig. 4. The frequency domain information is utilised. In this, the Frequency Domain Converter (FDC) is used to extract the amplitude and phase of the image. Furthermore, they are used to calculate the Manhattan distance between the blurred and sharp images in terms of amplitude and phase.

and phase distance $Loss_{FDC}$.

$$\begin{cases} gt_{amp}, gt_{pha} = FDC(gt_{img}) \\ pre_{amp}, pre_{pha} = FDC(pre_{img}) \\ Loss_{FDC} = L1(gt_{amp}, pre_{amp}) + L1(gt_{pha}, pre_{pha}). \end{cases} \tag{6}$$

3.4 Loss Function

The method proposed in this paper consists of two main losses: MSE and FDC losses. The entire loss function can be formulated as Eq. 7:

$$\mathcal{L} = \sigma_1 \mathcal{L}_{mse} + \sigma_2 \mathcal{L}_{fdc}, \tag{7}$$

where σ_1 and σ_2 are two weighting parameters, which are set to 1 and 0.1, respectively. The loss function $\mathcal{L}_{mse}$ is the mean-square error in pixels between the ground-truth and the prediction. $\mathcal{L}_{fdc}$ is the manhattan distance in the frequency domain between the ground-truth and the prediction.

4 Experimental Evaluation

4.1 Dataset

Beam-Splitter Deblurring Dataset (BSD) [24]. It uses a beam splitter acquisition system with two synchronized cameras and obtains sharp and blurred images by controlling the exposure time and exposure intensity. Unlike the way of obtaining blur from sharp frame degradation, BSD is a real-world dataset in the true sense of the word, containing three different frame rate datasets of 1ms-8ms, 2ms-16ms and 3ms-24ms. There are 60 training video datasets, 20 test video datasets, and 20 validation video datasets, each of which contains sharp frames and the corresponding blurred frames.

Gopro Dataset (Gopro) [25]. It is a dataset for deblurring tasks. The dataset consists of 3,214 blurred images of size 1,280 × 720, of which 2,103 are training images and 1,111 are test images. The dataset consists of one-to-one correspondence of real blurred images and ground-truth images, both captured by a high-speed camera.

REalistic and Diverse Scenes Dataset (Reds) [26]. It is a dataset that provides realistic and dynamic scenes for video deblurring and super-resolution. The dataset consists of 300 video sequences with a resolution of 720 × 1,280, of which 240 are training videos, 30 are validation videos, and 30 are test videos.

4.2 Experimental Setup

This paper was all performed on a Pytorch equipped with two 3090 GPUs. The code in this article is based on the ESTRNN [24]. The models are not pre-trained on the rest of the dataset. We train the model for 500 epochs with the Adam optimizer. The initial learning rate is 5×10^{-4}. We train the model using RGB patches of size 256 × 256 in subsequences of 5 frames as input. In addition, we implement horizontal and vertical flipping for each subsequence for data enhancement. The batch size is set to 8.

4.3 Ablation Study

We conducted ablation studies to demonstrate the effectiveness of DEF, CABs, and frequency domain information. As shown in Table 1 (a) and Table 1 (b), this paper conducts a series of comparison experiments to prove the effectiveness of the DEF. DEF has a large improvement over both RNN and Bi-LSTM.

In this paper, we set up a comparison where one group uses DABs to replace the GSA network, and the other group uses the Global Max Pooling(GMP) Operation to replace the DA in DABs. The results are shown in Table 1 (c), which shows that the structure of multi-branching can lead to an improvement of the network effect. And the network can have a more significant improvement

Table 1. Ablation study table. (a) is a comparison of the backbone network of the DEF(Ours) proposed in this paper with the RNN in ESTRNN. (b) is a comparison of the backbone network of the DEF and Bi-LSTM. (c) is a comparison of the different feature fusion networks. (d) is the effect of the use of information in the frequency domain with or without the use of the information in the frequency domain. For quantitative experimental comparisons, the remaining parameters are kept constant while comparing a particular parameter. The ablation studies are all performed on BSD for 2ms-16ms. (✗ denotes not using frequency information, ✓ denotes using frequency information. Best and second best scores are **highlighted** and underlined).

Group	Backbone	Frequency domain	Fusion network	PSNR↑	SSIM↑
(a)	RNN	✗	GSA	31.93	0.925
	DEF	✗	GSA	**32.61**	**0.932**
(b)	Bi-LSTM	✓	GSA	32.63	0.931
	DEF	✓	GSA	**32.98**	**0.937**
(c)	DEF	✓	GSA	32.98	0.937
	DEF	✓	GMP	32.99	**0.938**
	DEF	✓	DABs(Ours)	**33.03**	0.938
(d)	DEF	✗	GSA	32.61	0.932
	DEF	✓	GSA	**32.98**	**0.937**

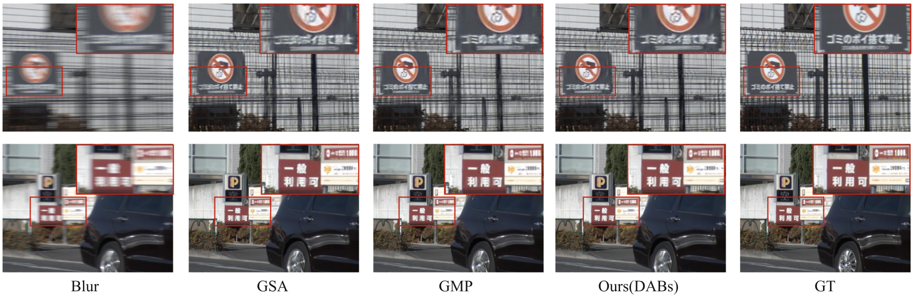


Fig. 5. Visualization of different feature fusion networks.

only when DA is used instead of GMP. Moreover, analyzing the visual effect in Fig. 5, it can be found that the left part of the text produced by using the DABs module is clearer and has fewer color differences. The overall visual effect of DABs is better than the results of the other two modules.

The effect of DABs is better than GSA and GMP because the GSA structure focuses more on the result between two frames after various transformations. Moreover, the transformations contain much content, and learning the difference between the two frames requires more time. DABs can speed up the convergence process while focusing on the difference between frames, as shown in Fig. 6.

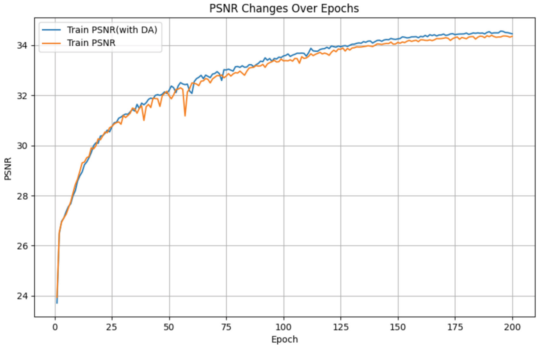


Fig. 6. Plot of DA Block's impact on PSNR metrics during training. The model converges better and is more stable when using the DA Block (blue line). (Color figure online)

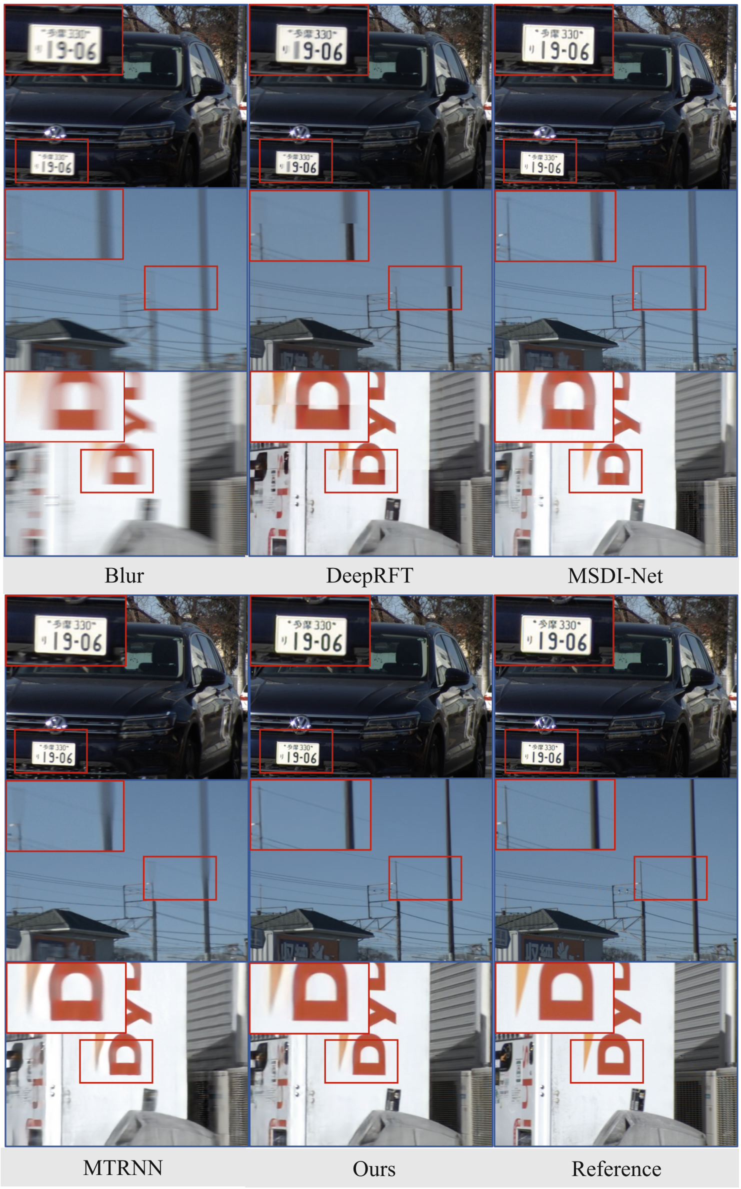
Finally, based on the ablation experiments, we find that the introduction of frequency domain information significantly improves network performance, as shown in Table 1 (d).

4.4 Quantitative and Visual Results

Table 2. Quantitative results of our method comparing other methods on the BSD dataset(Best and second best scores are **highlighted** and underlined).

Methods	Reference	1ms8ms		2ms16ms		3ms24ms	
		PSNR↑	SSIM↑	PSNR↑	SSIM↑	PSNR↑	SSIM↑
STRCNN	(ICCV 2017) [11]	32.20	0.924	30.33	0.902	29.42	0.893
DBN	(CVPR 2017) [7]	33.22	0.935	31.75	0.922	31.21	0.922
IFI-RNN	(CVPR 2017) [9]	33.00	0.933	31.53	0.919	30.89	0.917
SRN	(ICCV 2018) [27]	31.84	0.917	29.95	0.891	28.92	0.882
STFAN	(ICCV 2019) [10]	32.78	0.922	32.19	0.919	29.47	0.872
MTRNN	(ECCV 2020) [15]	28.06	0.868	26.85	0.841	27.17	0.866
CDVD-TSP	(CVPR 2020) [28]	33.54	**0.942**	32.16	0.926	31.58	0.926
MSDI-Net	(ECCV 2022) [29]	28.40	0.885	27.87	0.865	28.03	0.875
DeepRFT	(AAAI 2023) [30])	29.81	0.902	29.76	0.910	28.14	0.890
ESTRNN	(IJCV 2023) [24]	33.36	0.937	31.95	0.925	31.39	0.926
Ours		**33.68**	0.938	**33.03**	**0.938**	**31.66**	**0.929**

We conducted side-by-side comparison experiments on the real-world dataset BSD using many methods [7,9–11,24,27–30]. The visual results are shown in



Blur DeepRFT MSDI-Net

MTRNN Ours Reference

Fig. 7. Visual results of our method compared to other methods.

Fig. 7. Our approach achieves more visually appealing results in BSD. The quantitative results are shown in Table 2. Compared to all methods, our method reaches the state-of-the-art. Both visual and quantitative results validate the effectiveness of our method in real-world video deblurring tasks.

In addition, our method conducts comparative experiments on the Gopro and Reds datasets, and the quantitative results are shown in Table 3. Our methods also reach state-of-the-art.

Table 3. Quantitative results of our method comparing other methods on the Gopro and Reds datasets(Best and second best scores are **highlighted** and underlined).

Method	Reference	Gopro		Reds	
		PSNR↑	SSIM↑	PSNR↑	SSIM↑
STRCNN	(ICCV 2017) [11]	28.74	0.8465	30.23	0.8708
DBN	(CVPR 2017) [7]	29.91	0.8823	31.55	0.8960
IFI-RNN(c2h1)	(CVPR 2017) [9]	29.79	0.8817	31.29	0.8913
IFI-RNN(c2h2)	(CVPR 2017) [9]	29.92	0.8838	31.35	0.8929
IFI-RNN(c2h3)	(CVPR 2017) [9]	29.97	0.8859	31.36	0.8942
STFAN	(ICCV 2019) [10]	30.51	0.9054	32.03	0.9024
ESTRNN	(IJCV 2023) [24]	**31.07**	0.9023	32.63	0.9110
Ours		31.01	**0.9131**	**33.22**	**0.9240**

5 Conclusion

This paper proposes an efficient video deblurring method (STFFNet) for perceptually oriented and metrically favourable enhancement. Specifically, we first explore using an Encoder-Decoder to construct a novel backbone. It combines global features while generating current frame features to extract more profound and broader information for better recovery. In addition, we develop a new feature fusion module to speed up the fitting and improve the modelling results. Finally, we added frequency domain information to the network to make the network more focused on high-frequency information, resulting in more explicit images. Experimental results show that STFFNet performs up to the current state-of-the-art methods in the benchmark datasets.

Acknowledgements. This work is partly supported by the Chongqing University of Technology high-quality development Action Plan for of graduate education (gzlcx20233188), the Chongqing Postgraduate Research and Innovation Project Funding (No. CYS23677) and the Youth Project of Science and Technology Research Program of Chongqing Education Commission of China (No. KJQN202401106).

References

1. Chan, K.C., Zhou, S., Xu, X., Loy, C.C.: BasicVSR++: improving video super-resolution with enhanced propagation and alignment. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 5972–5981 (2022)

2. Haris, M., Shakhnarovich, G., Ukita, N.: Space-time-aware multi-resolution video enhancement. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 2859–2868 (2020)
3. Liu, X., Kong, L., Zhou, Y., Zhao, J., Chen, J.: End-to-end trainable video super-resolution based on a new mechanism for implicit motion estimation and compensation. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pp. 2416–2425 (2020)
4. Shi, Z., Liu, X., Shi, K., Dai, L., Chen, J.: Video frame interpolation via generalized deformable convolution. IEEE Trans. Multimedia **24**, 426–439 (2021)
5. Shi, Z., Xu, X., Liu, X., Chen, J., Yang, M.H.: Video frame interpolation transformer. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 17482–17491 (2022)
6. Zhang, F., Li, Y., You, S., Fu, Y.: Learning temporal consistency for low light video enhancement from single images. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 4967–4976 (2021)
7. Su, S., Delbracio, M., Wang, J., Sapiro, G., Heidrich, W., Wang, O.: Deep video deblurring for hand-held cameras. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1279–1288 (2017)
8. Wang, X., Chan, K.C., Yu, K., Dong, C., Change Loy, C.: EDVR: video restoration with enhanced deformable convolutional networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (2019)
9. Nah, S., Son, S., Lee, K.M.: Recurrent neural networks with intra-frame iterations for video deblurring. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 8102–8111 (2019)
10. Zhou, S., Zhang, J., Pan, J., Xie, H., Zuo, W., Ren, J.: Spatio-temporal filter adaptive network for video deblurring. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 2482–2491 (2019)
11. Hyun Kim, T., Mu Lee, K., Scholkopf, B., Hirsch, M.: Online video deblurring via dynamic temporal blending network. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 4038–4047 (2017)
12. Wieschollek, P., Hirsch, M., Scholkopf, B., Lensch, H.: Learning blind motion deblurring. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 231–240 (2017)
13. Du, B., Ren, X., Ren, J.: CNN-based image super-resolution and deblurring. In: Proceedings of the 2019 International Conference on Video, Signal and Image Processing, pp. 70–74 (2019)
14. Cai, J., Zuo, W., Zhang, L.: Dark and bright channel prior embedded network for dynamic scene deblurring. IEEE Trans. Image Process. **29**, 6885–6897 (2020)
15. Zhang, X., Wang, T., Jiang, R., Zhao, L., Xu, Y.: Multi-attention convolutional neural network for video deblurring. IEEE Trans. Circuits Syst. Video Technol. **32**(4), 1986–1997 (2021)
16. Ren, W., et al.: Deblurring dynamic scenes via spatially varying recurrent neural networks. IEEE Trans. Pattern Anal. Mach. Intell. **44**(8), 3974–3987 (2021)
17. Zhu, C., et al.: Deep recurrent neural network with multi-scale bi-directional propagation for video deblurring. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 36, pp. 3598–3607 (2022)
18. Huang, Z., Xu, W., Yu, K.: Bidirectional LSTM-CRF models for sequence tagging. arXiv preprint arXiv:1508.01991 (2015)
19. Cho, K., et al.: Learning phrase representations using RNN encoder-decoder for statistical machine translation. arXiv preprint arXiv:1406.1078 (2014)

20. Sutskever, I., Vinyals, O., Le, Q.V.: Sequence to sequence learning with neural networks. In: Advances in Neural Information Processing Systems, vol. 27 (2014)
21. Cai, M., Zhang, H., Huang, H., Geng, Q., Li, Y., Huang, G.: Frequency domain image translation: more photo-realistic, better identity-preserving. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 13930–13940 (2021)
22. Jiang, J., et al.: Five a+ network: you only need 9k parameters for underwater image enhancement. arXiv preprint arXiv:2305.08824 (2023)
23. Cooley, J.W., Tukey, J.W.: An algorithm for the machine calculation of complex fourier series. Math. Comput. **19**(90), 297–301 (1965)
24. 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)
25. Nah, S., Hyun Kim, T., Mu Lee, K.: Deep multi-scale convolutional neural network for dynamic scene deblurring. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3883–3891 (2017)
26. Nah, S., et al.: NTIRE 2019 challenge on video deblurring and super-resolution: Dataset and study. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (2019)
27. Tao, X., Gao, H., Shen, X., Wang, J., Jia, J.: Scale-recurrent network for deep image deblurring. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 8174–8182 (2018)
28. Pan, J., Bai, H., Tang, J.: Cascaded deep video deblurring using temporal sharpness prior. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 3043–3051 (2020)
29. Li, D., Zhang, Y., Cheung, K.C., Wang, X., Qin, H., Li, H.: Learning degradation representations for image deblurring. In: European Conference on Computer Vision, pp. 736–753. Springer (2022)
30. Mao, X., Liu, Y., Liu, F., Li, Q., Shen, W., Wang, Y.: Intriguing findings of frequency selection for image deblurring. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 37, pp. 1905–1913 (2023)

Self-distilled Dual-Network with Pixel Screening Loss for Blind Image Deblurring

Tianyi Li, Ming Tian, Changxin Gao(✉), and Nong Sang

National Key Laboratory of Multispectral Information Intelligent Processing Technology, School of Artificial Intelligence and Automation, Huazhong University of Science and Technology,Wuhan, China
{litianyi123,tianming,cgao,nsang}@hust.edu.cn

Abstract. In the field of image processing, blind image deblurring aims to restore sharp details in images blurred by an unknown convolution kernel. Recent advancements have shown that deep networks can act as effective image generative priors (DIP) for restoring clear images without requiring external datasets. However, the inherent non-uniqueness of solutions in blind image deblurring often leads DIP-based methods to converge on local optima, resulting in over- or under-deblurred images. To overcome this limitation, we propose a novel deblurring framework featuring dual image generators. These generators mutually constrain each other during training, guiding the model towards the optimal solution. Building on our network structure, we employ a self-ensemble and self-distillation strategy to guide network training, further enhancing performance. Additionally, we introduce a novel loss function based on a pixel screening method, which focuses on the important pixels. This loss enables the network to model the blur kernel more accurately and facilitates the restoration of image details. Our experiments demonstrate that our deblurring approach outperforms most existing methods both qualitatively and quantitatively.

Keywords: Blind Image Deblurring · Dual-Network · Self-Distillation · Pixel Screening

1 Introduction

Blind image deblurring (BID) is a classical problem in the domain of image restoration, focusing on the elimination of unknown blur caused by camera shake. When the blur kernel is spatially invariant, the blurred image B to be restored can be formulated as:

$$B = I \otimes k + n, \tag{1}$$

where I represents the sharp image we aim to recover, $\otimes$ denotes the 2D convolution operation, k is the blur kernel, and n represents the additive white Gaussian

A. Antonacopoulos et al. (Eds.): ICPR 2024, LNCS 15332, pp. 167–181, 2025.
https://doi.org/10.1007/978-3-031-78125-4_12

noise. Thus, in the task of blind image deblurring, our goal is to estimate both I and k from the blurred image B. Given the multitude of solutions for I and k that satisfy the equation, this problem is also a typical example of an ill-posed problem.

Most of the existing traditional optimization-based methods for addressing this issue are based on the Maximum a Posteriori (MAP) framework,

$$(k, I) = \arg\max_{k,I} P(k, I|B) = \arg\max_{k,I} P(B|k, I)P(I)P(k), \tag{2}$$

where $P(B|k, I)$ is the likelihood term and $P(I)$, $P(k)$ model the priors of clean image and blur kernel, respectively. Numerous outstanding works focus on the design of priors $P(I)$ and $P(k)$ in (2), to enhance the accuracy of kernel estimation [3,7,18,24,41]. Although these manually designed priors are powerful, they still have limitations in fully modeling both the sharp images and the blur kernels.

Ulyanov et al. [33] demonstrated that neural network architectures possess an intrinsic image generative prior, termed as "Deep Image Prior" (DIP), which can be leveraged for image restoration tasks. This finding suggests that Convolutional Neural Networks (CNNs) can inherently grasp low-level statistical information [20], acting as an effective prior without the need for training on extensive datasets. Building on the concept of DIP, Ren et al. [26] proposed using a DIP network, specifically the asymmetric autoencoder [28] with skip connections [33], along with a fully connected network (FCN) to map noise inputs to sharp images and blur kernels. The optimization problem for this framework can be formulated as:

$$\min_{G_I, G_k} \|G_I\left(Z_I\right) \otimes G_k\left(Z_k\right) - B\|^2, \tag{3}$$

where G_I and G_k are the generators, Z_I and Z_k are noise inputs.

Motivation. Recently, several efforts have been made to optimize this framework [1,4,9,17,32]. For instance, Tian et al. [32] proposed using CNNs to model the blur kernel and incorporating attention mechanisms into the image generator to better capture priors, Bredell et al. [1] suggested using Wiener deconvolution to guide DIP during optimization and achieved more stable deblurring performance. However, due to the non-uniqueness of solutions, DIP-based methods are prone to converging to local optima, making it challenging to accurately restore image details. Additionally, since the image and kernel generators are optimized simultaneously, an erroneous solution generated by any one network can affect the final results. Therefore, we aim to develop a new deblurring method to address the inherent issues of DIP-based blind deblurring approaches.

Our Approach. In this paper, we introduce a network framework with dual image generators (See § 3.1). These two generators work simultaneously to model the image, where their mutual constraints help prevent the model from converging to local optima, significantly improving image detail restoration. We also utilize a self-ensemble and self-distillation strategy that ensembles the clear image outputs of the network and incorporates them into the loss function to guide network training, further enhancing performance (See § 3.2). Inspired by recent work [40], which introduced a pixel screening method to exclude the impact of adverse pixels, we propose a novel loss function (See § 3.3). This loss function weights the importance of each pixel in loss calculation, allowing the network to focus more on the parts beneficial for modeling the blur kernel during training. A more precise blur kernel helps the image generators produce images with finer details, thereby enhancing the overall performance of the network.

Contributions. Our contributions in this paper are as follows:

- We propose a network architecture with dual image generators that effectively mitigates the model's convergence to local optima. Building upon our proposed model, we further employ a self-ensemble and self-distillation strategy to enhance the network's performance.
- We propose a pixel screening loss, aiding the network in more accurately modeling the blur kernel, thereby enhancing the restoration of image details.
- Qualitative and quantitative experiments on popular datasets demonstrate that our method is competitive with state-of-the-art approaches.

2 Related Work

2.1 Optimization-Based Deblurring

A popular traditional approach for blind image deblurring is based on the Maximum a Posteriori (MAP) framework. Research in this area primarily focuses on designing prior constraints for sharp images and blur kernels. These methods aim to make intermediate images closer to the clear image and improve the accuracy of blur kernel estimation. Notable priors include total variation (TV) [3], l_0-norm gradient prior [38], edge-based patch priors [30], and color-line prior [13].

On top of or instead of the aforementioned priors, recent works have proposed stronger ones, such as dark channel prior [24], local minimal intensity prior [35], and superpixel segmentation prior [19], achieving state-of-the-art results. However, despite the effectiveness of these manually designed priors, they still face limitations in fully representing both sharp images and blur kernels.

2.2 Supervised Learning-Based Deblurring

Many existing deep learning methods address this by constructing datasets with clear images and true kernels to train neural networks in a supervised manner. Some studies [2,29] use deep neural networks to directly learn map blurry images to the blur kernels that need to be estimated. However, due to the diversity of the motion blur kernels, it is impossible to construct a comprehensive training set that enables the network to estimate arbitrary blur kernels.

Therefore, many methods focus on kernel-free neural networks, which aim to directly map blurry images to their corresponding sharp images. Nah et al. [22] use a "multi-scale" training strategy to train a deep neural network composed of stacked residual blocks. Tao et al. [31] build on this by introducing inter-scale weight sharing, which reduces the number of parameters in the network. Zhang et al. [39] employ a recursive weight setting method to increase the receptive field of the network, allowing it to handle more severe motion blur. Kupyn et al. [12] use generative adversarial networks (GANs) to obtain more realistic sharp images. While these learning-based deblurring methods can effectively address non-uniform blur issues, their performance is limited when dealing with severe blur and is constrained by the datasets.

2.3 DIP-Based Deblurring

To avoid the collection cost and generalization issues associated with external training sets in deep learning, recent studies [1,4,9,17,26,32] focus on developing dataset-free blind image deblurring methods. These methods are based on the Deep Image Prior (DIP) [33], which suggests that convolutional neural networks can act as implicit regularizers when fitting images using random seeds as input.

Inspired by this, Ren et al. [26] proposed SelfDeblur to address the blind deblurring problem. This method uses a CNN to fit the image while employing a fully connected network to model the prior of the blur kernel. Jan et al. [9] incorporated ideas from traditional MAP methods into SelfDeblur, enhancing the stability and performance of the deblurring process. Tian et al. [32] suggested using CNNs to model the blur kernel and integrating attention mechanisms into the image generator to more effectively capture the priors. However, the problem of solution non-uniqueness in blind image deblurring can cause DIP-based methods to converge on local optima, leading to images that are either over-deblurred or under-deblurred. Our method is also based on DIP, but we aim to alleviate the issue of neural networks converging to local optima when fitting images. Additionally, we seek to better guide network training to achieve more accurate results.

3 Method

3.1 Dual-Network Architecture

Most DIP-based methods focus on using one network to model the sharp image and another network to model the blur kernel separately. However, since the blind image deblurring problem is a typical ill-posed problem, it is highly prone to converging to suboptimal solutions, resulting in unsatisfactory sharp images.

Hence, we propose a novel dual-network architecture, which is shown in Fig. 1. In our framework, there are two DIP networks dedicated to modeling the sharp image. These networks independently map their respective noise inputs to clear images. Consequently, our optimization problem is defined as follows:

$$\min_{G_I^1, G_I^2, G_k} L_{data}\left(G_I^1\left(Z_I^1\right) \otimes G_k\left(Z_k\right); B\right) + L_{data}\left(G_I^2\left(Z_I^2\right) \otimes G_k\left(Z_k\right); B\right), \quad (4)$$

where L_{data} is the data-term measuring the degradation model error. This error can be measured by its l^2 norm as in (3) or others. Given that the inputs of the two noises are different, the image generation networks G_I^1 and G_I^2 will learn distinct parameter distributions, which in turn leads to varied output results. Nonetheless, as the images generated by both networks are convolved with the same blur kernel, they independently aid G_k in accurately modeling the blur kernel. Throughout the optimization process, the dual image generators mutually constrain each other, effectively preventing the model from converging to suboptimal solutions.

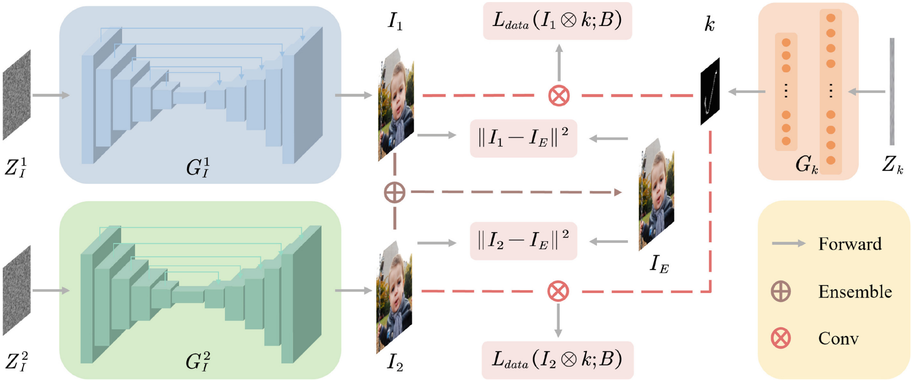


Fig. 1. Overview of proposed method. The dual image generators G_I^1 and G_I^2 estimate sharp images, while G_k generates the blur kernel. The error with the blurred image B is measured using L_{data} which can be either L_{ps} (See §3.3) or SSIM, and the ensemble output is utilized to guide the network training process.

3.2 Self-ensemble and Self-distillation

Chen et al. [4] have demonstrated that ensembling deblurred images can significantly enhance the visual quality of images. However, they employed meticulously designed complex modules to ensure image alignment. Benefiting from our dual-network architecture, we can also average the two clear images generated by the dual image generators to achieve better results. Unlike training two image generators independently, our image generators are simultaneously constrained by the same blur kernel. As a result, the sharp images we obtain are naturally aligned, allowing us to directly integrate the two images without additional alignment operations.

Additionally, to achieve better integration results, we aim to combine the outputs of the image generator at each step of the iterative process. Therefore, we have adopted the concept of temporal ensembling [15] by calculating the Exponential Moving Average (EMA) of the results from each image generator and then averaging these to derive the final integrated result.

Let I_1, I_2 respectively represent the $G_I^1\left(Z_I^1\right)$ and $G_I^2\left(Z_I^2\right)$ at each step of the training process, the EMA output for each generator $\widetilde{I}_i$ $(i = 1, 2)$ at each step can be computed by:

$$\widetilde{I}_i \leftarrow \mu \widetilde{I}_i + (1 - \mu) I_i, \tag{5}$$

where μ is a hyperparameter that takes values ranging from zero to one. Their total integrated output I_E can be represented as:

$$I_E = \frac{1}{2}\left(\widetilde{I}_1 + \widetilde{I}_2\right). \tag{6}$$

While self-ensemble can enhance performance, the base model itself remains unaffected by the integrated result. To address this, we further apply knowledge distillation [6] to improve the base model's capabilities. The self-distillation loss function can be written as:

$$L_{sd} = \|I_1 - I_E\|^2 + \|I_2 - I_E\|^2. \tag{7}$$

This means that at each iteration step, the superior integrated output guides the network training, allowing the model to converge towards a better solution. Our ablation studies show that with the help of self-ensemble and self-distillation, the model tends to exhibit improved performance. This technique effectively enhances the network's ability to learn from its own predictions, thereby refining its output and contributing to overall better results.

3.3 Pixel Screening Loss

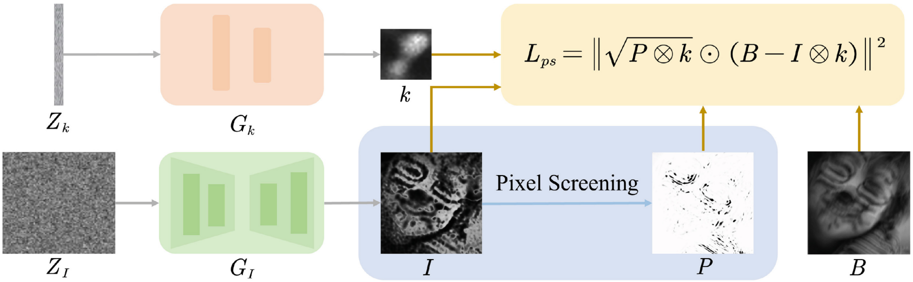

$L_{ps} = \left\|\sqrt{P \otimes k} \odot (B - I \otimes k)\right\|^2$
Z_k
G_k
k
Pixel Screening
Z_I
G_I
I
P
B

Fig. 2. Pixel screening loss. This involves calculating the pixel screening map of the intermediate generated image and using it as a weight combined with the Mean Squared Error (MSE).

In traditional optimization-based methods of blind image deblurring [23,24,27], the blur kernel and intermediate image are typically estimated alternately, ultimately aiming to converge to the blur kernel. Zhang et al. [40] have proposed an intermediate image correction method that leverages Bayesian posterior estimation. This method screens through the intermediate image to identify and exclude unfavorable pixels, thereby reducing their impact on kernel estimation.

As described by Zhang et al. [40], given the estimated I and k, the pixel screening map P_{ij} can be explicitly represented as:

$$P_{ij} = \frac{\mathcal{N}\left((I \otimes k)_{ij}, \sigma^2\right)(1 - P_0)}{\mathcal{N}\left((I \otimes k)_{ij}, \sigma^2\right)(1 - P_0) + cP_0}, \tag{8}$$

where P_0 is a parameter that can be adjusted and $c = \frac{1}{c_{\max} - c_{\min}}$, $c_{\max}$ and $c_{\min}$ are for the range of the image values. However, this method is primarily applicable to traditional optimization-based approaches. Since neural network methods estimate both the image and the blur kernel simultaneously and involve gradient backpropagation, the intermediate image correction method cannot be directly applied.

To integrate this approach with neural network methods, we propose a novel loss function, which can be written as:

$$L_{ps} = \|\sqrt{P \otimes k} \odot (B - I \otimes k)\|^2, \tag{9}$$

where I denotes $G_I^1\left(Z_I^1\right)$ or $G_I^2\left(Z_I^2\right)$, k denotes $G_k\left(Z_k\right)$, $\odot$ is the pointwise multiplication operator, and P represents the pixel screening map applied to I. This map, P, is calculated using the intermediate results I and k obtained during the network's training process. The loss function underscores the significance of each pixel's contribution to the loss by performing a weighted average, based on the correctness of the contribution, rather than a simple average. This specialized loss helps the model more accurately model the blur kernel, leading to improved performance. The pipeline of this loss function is shown in Fig. 2.

3.4 Total Loss

In addition to the previously mentioned loss, we also use l^2 regularization for the kernel values, and the the data-term is switched from Lps to the Structural Similarity Index Measure (SSIM) after a certain number of steps. The total loss for training our model can be represented as:

$$L_{total} = L_{data}\left(I_1 \otimes k; B\right) + L_{data}\left(I_2 \otimes k; B\right) + \lambda L_{sd} + \beta\|k\|^2, \tag{10}$$

where L_{data} is L_{ps} if iter < 2000, SSIM otherwise, $\lambda = 1$ if iter > 2000, 0 otherwise, and $\beta = 0.1$.

4 Experiments

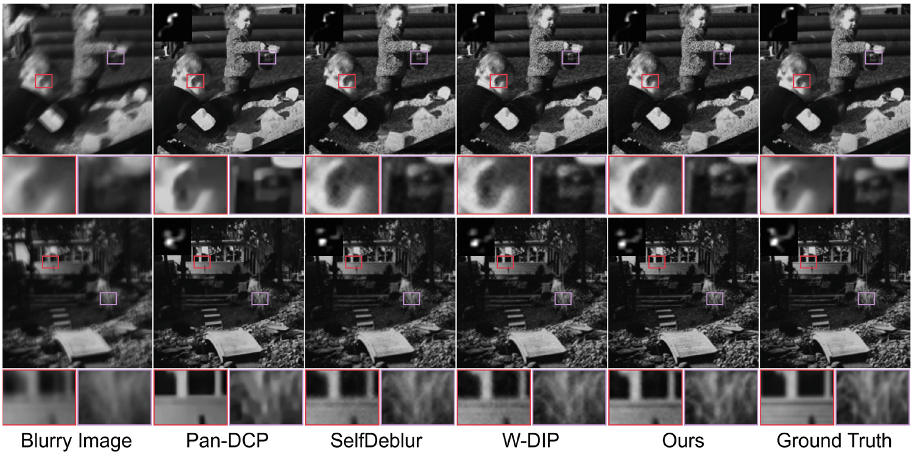


Fig. 3. Visual comparison on sample images from Levin et al. [16]. The estimated kernels are displayed in the top left corner of the image.

Table 1. Average PSNR, SSIM comparison on the dataset of Levin et al., $^{\triangle}$ indicates the final deblurring results are using the method from [34].

Method	PSNR	SSIM
Krishnan et al.$^{\triangle}$ [11]	29.88	0.8666
Cho&Lee$^{\triangle}$ [5]	30.57	0.8966
Levin et al.$^{\triangle}$ [34]	30.80	0.9092
Xu&Jia$^{\triangle}$ [37]	31.67	0.9163
Sun et al.$^{\triangle}$ [30]	32.99	0.9330
Zuo et al.$^{\triangle}$ [41]	32.66	0.9332
Pan-DCP$^{\triangle}$ [24]	32.69	0.9284
SRN [31]	23.43	0.7117
SelfDeblur [26]	33.07	0.9313
W-DIP [1]	33.61	0.9329
Ours	**34.79**	**0.9446**

4.1 Implementation Details

The method we proposed is implemented in PyTorch. We use the Adam [8] optimizer to adjust the parameters of both the image generator and the blur kernel generator, and each employs a different learning rate. We set the initial learning rate to $1e-4$ for image generators and $1e-2$ for the kernel generator, and decrease it by multiplying with 0.5 at the 2K, 3K, and 4K iterations following the implementation by Ren et al. [26]. For the evalutions, we run this optimization process for 5K iterations, and after 2K iterations, we apply the EMA (Exponential Moving Average) algorithm with a parameter set to 0.9. Additionally, for the pixel screening loss, we set the hyperparameter P_0 to 0.1. It is also worth emphasizing that, similar to other DIP-based methods, our approach is self-supervised and does not require any external training dataset.

4.2 Comparison with the State-of-the-Art Methods

We opt to conduct our experiments using the popular datasets from Levin et al. [16] and Lai et al. [14]. We use the ensemble output of the method we proposed as final results. For traditional optimization-based approaches, we begin by estimating the blur kernel using these methods, followed by employing a non-blind deconvolution method to recover the sharp image.

Results on the Dataset of Levin et al. On the Levin dataset, we compared several traditional optimization-based methods and neural network methods based on DIP [33], among which we reproduced the results of method [1] using its source code. The results in Table. 1 indicate that our method outperforms others in terms of both PSNR and SSIM, achieving the best results. At the same time, it can be seen from Fig. 3 that, compared to other methods, our method can obtain more image details and a more accurate blur kernel.

Results on the Dataset of Lai et al. While the dataset of Levin et al. [16] consists of 32 greyscale images with relatively small kernels, the Lai et al. [14] dataset includes 25 clean color images and 4 large blur kernels. The blurry images in the Lai et al. dataset are organized into five categories: Manmade, Natural, People, Saturated, and Text. Each category contains 20 blurry images, providing a diverse set of images for comprehensive evaluation. We present our results alongside those of other methods in Table. 2, demonstrating that our method remains competitive even on more challenging datasets.

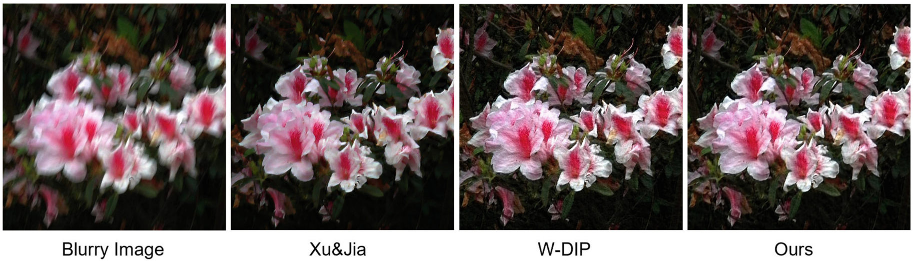


Fig. 4. Real-world blur image from the Lai et al. [14] dataset comparing our method with other methods.

Table 2. Average PSNR/SSIM comparison of 5 categories on the dataset of Lai et al. [14]. The methods marked with $^\triangle$ adopt [10] and [36] as non-blind deconvolution after blur kernel estimation.

Images	Cho&Lee$^\triangle$ [5]	Xu&Jia$^\triangle$ [37]	Michaeli et al. $^\triangle$ [21]	Perrone et al.$^\triangle$ [25]
Manmade	16.35/0.3890	19.23/0.6540	17.43/0.4189	17.41/0.5507
Natural	20.14/0.5198	23.03/0.7542	20.70/0.5116	21.04/0.6764
People	19.90/0.5560	25.32/0.8517	23.35/0.6999	22.77/0.7347
Saturated	14.05/0.4927	14.79/0.5632	14.14/0.4914	14.24/0.5107
Text	14.87/0.4429	18.56/0.7171	16.23/0.4686	16.94/0.5927
Average	17.06/0.4801	20.18/0.7080	18.37/0.5181	18.48/0.6130
Images	Pan-DCP$^\triangle$ [24]	SelfDeblur [26]	W-DIP [1]	Ours
Manmade	18.59/0.5942	20.35/0.7543	20.23/0.7406	**20.43/0.7713**
Natural	**22.60**/0.6984	22.05/0.7092	22.30/0.7382	22.49/**0.7525**
People	24.03/0.7719	25.94/0.8834	**26.29/0.8991**	26.18/0.8932
Saturated	16.52/0.6322	16.35/0.6364	17.06/0.6717	**17.12/0.6847**
Text	17.42/0.6193	20.16/0.7785	20.25/0.7719	**20.93/0.7928**
Average	19.89/0.6656	20.97/0.7524	21.23/0.7643	**21.43/0.7789**

Results on the Real-World Images. We also test our method on real-world blurry images, which typically involve unknown blur kernels and pose significant challenges. As shown in Fig. 4, our method produces results that are as good as or even better than other blind image deblurring methods.

4.3 Ablation Study

Network Architecture. To evaluate the benefits of our proposed dual-generator network structure, we conducted experiments on the Levin dataset three times and recorded the distribution of all results based on the PSNR metric. We did not use any additional proposed losses and focused solely on the impact of the network structure. As shown in Fig. 5, with all other factors being the same, the dual-generator structure resulted in overall better performance compared to a single generator. We also observed that, although the images generated by the two generators had a similar overall distribution, they still exhibited differences. This indicates that the two generators can function independently while also constraining each other during training, helping the model avoid converging to suboptimal solutions.

We further designed experiments to investigate the impact of different numbers of generators (ranging from 1 to 4) on network performance. This evaluation was also done without incorporating the other proposed losses, focusing solely on the effect of changing the quantity of generators. As indicated in the results presented in Table. 3, we observe a gradual improvement in performance with an increasing number of generators. The most notable improvement is observed

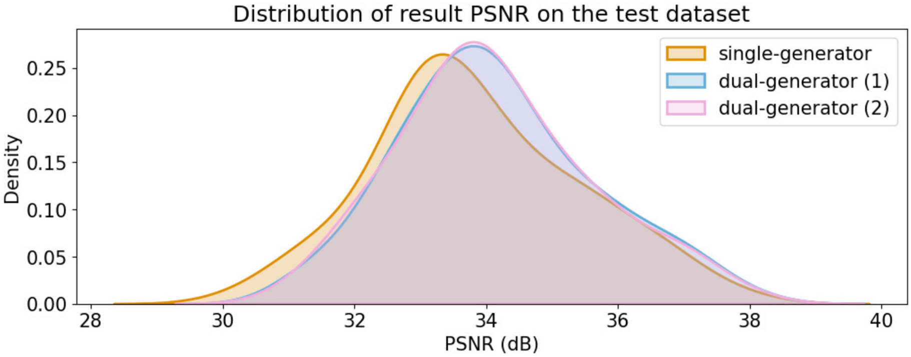


Fig. 5. The PSNR results distribution for whether using the dual-generator structure on the Levin et al. [16].

Table 3. PSNR results of our method with different quantities of image generators, tested on Levin et al. [16].

generator numbers	1	2	3	4
PSNR	34.01	34.61	34.74	34.80

when the number of generators is set to two. Taking into account the balance between model size and performance, we opt for two image generators for our network structure, as this setup provides a significant enhancement in performance without excessively increasing the complexity of the model.

Self-ensemble. Figure 6 shows the PSNR results of our method on each image in the Levin dataset, highlighting both the outputs of the individual image generators and the results integrated using our self-ensemble method. As can be seen, integration improves performance for almost every image. This indicates that using EMA and integrating the outputs of the two generators can yield more accurate sharp images, laying the groundwork for our subsequent use of self-distillation.

Self-distillation and Pixel Screening Loss. Table 4 highlights the improvements in the PSNR metric for the model when utilizing self-distillation and pixel screening loss. In this context, "Max" refers to using the maximum metric value from the outputs of the two generators as the final result, whereas "Ensemble" indicates the use of the integrated output from both generators.

The data reveal that implementing the self-distillation loss improves PSNR values for both individual outputs and the integrated output. Additionally, the inclusion of pixel screening loss further optimizes the model's performance, achieving the highest improvements. Figure 7 shows a visual quality comparison with and without using pixel screening loss. This demonstrates the effectiveness

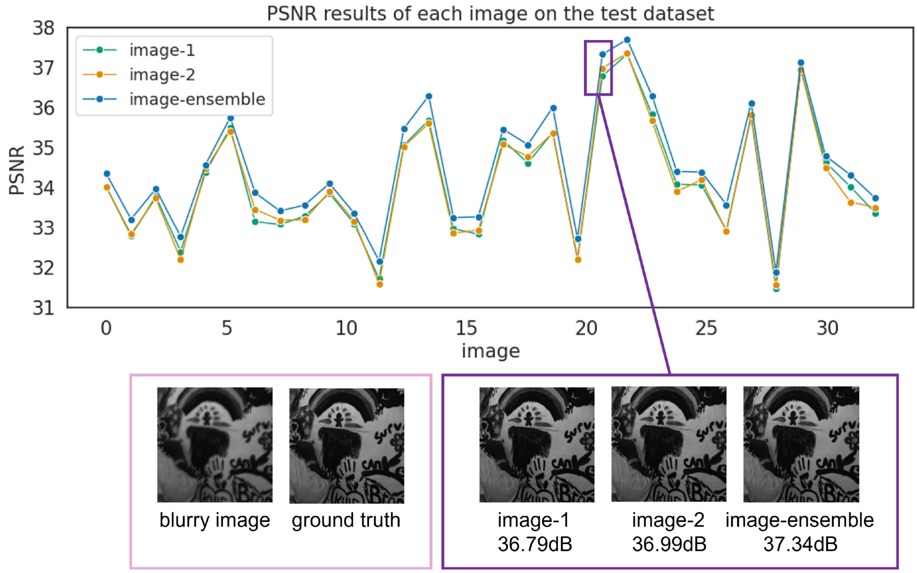
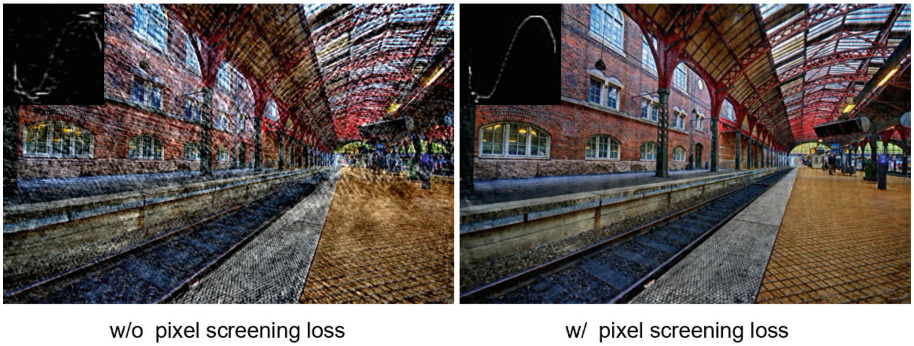


Fig. 6. The PSNR results of each generator's output and the self-ensemble output for each image in the dataset.


w/o pixel screening loss

w/ pixel screening loss

Fig. 7. Visual quality comparison of whether using pixel screening loss on an image in the dataset of Lai et al. [14]. The estimated kernels are displayed in the top left corner of the image.

Table 4. Ablation study of the proposed method on Levin et al.

Method	Max	Ensemble
Dual-Network	34.26	34.61
Dual-Network + Self-Distillation	34.49	34.68
Dual-Network + Self-Distillation + Pixel Screening Loss	34.56	34.79

of combining these techniques: self-distillation maximizes the benefits of integrating outputs from multiple generators, while pixel screening loss refines the focus on more accurate pixel contributions. Together, these techniques drive the model toward optimal performance.

5 Conclusions

In this paper, we propose a novel dual-network architecture featuring two image generators, complemented by self-distillation and pixel screening loss, aimed at tackling the challenge of blind image deblurring. Through comprehensive experiments, we demonstrate that our approach effectively improves the restoration of sharp image details by mitigating the model's convergence to suboptimal solutions and refining the accuracy of blur kernel estimation. The results confirm that our method not only enhances the visual quality of deblurred images but also outperforms existing methods, marking a significant advancement in the field of image processing. In the future, we will explore how to apply our method to other image restoration problems, such as non-uniform blind image deblurring.

References

1. Bredell, G., Erdil, E., Weber, B., Konukoglu, E.: Wiener guided dip for unsupervised blind image deconvolution. In: Proceedings of the IEEE/CVF Winter Conference Applications Computer Vision, pp. 3047–3056 (2023)
2. Chakrabarti, A.: A neural approach to blind motion deblurring. In: Proceedings of the European Conference Computer Vision, pp. 221–235. Springer (2016)
3. Chan, T.F., Wong, C.K.: Total variation blind deconvolution. IEEE Trans. Image Process. **7**(3), 370–375 (1998)
4. Chen, M., Quan, Y., Xu, Y., Ji, H.: Self-supervised blind image deconvolution via deep generative ensemble learning. IEEE Trans. Circuit Syst. Video Technol. **33**(2), 634–647 (2022)
5. Cho, S., Lee, S.: Fast motion deblurring. ACM Trans. Graph. **28**(5), 1–8 (2009)
6. Hinton, G., Vinyals, O., Dean, J.: Distilling the knowledge in a neural network. arXiv preprint arXiv:1503.02531 (2015)
7. Joshi, N., Szeliski, R., Kriegman, D.J.: PSF estimation using sharp edge prediction. In: Proceedings of the IEEE Conference Computer Vision Pattern Recognition, pp. 1–8 (2008)
8. Kingma, D.P., Ba, J.: Adam: a method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
9. Kotera, J., Šroubek, F., Šmídl, V.: Improving neural blind deconvolution. In: Proceedings of the IEEE International Conference Image Processing, pp. 1954–1958 (2021)
10. Krishnan, D., Fergus, R.: Fast image deconvolution using hyper-Laplacian priors. In: Proceedings of the NeurIPS, pp. 1033–1041 (2009)
11. Krishnan, D., Tay, T., Fergus, R.: Blind deconvolution using a normalized sparsity measure. In: Proceedings of the IEEE Conference Computer Vision Pattern Recognition, pp. 233–240 (2011)

12. Kupyn, O., Budzan, V., Mykhailych, M., Mishkin, D., Matas, J.: DeblurGAN: blind motion deblurring using conditional adversarial networks. In: Proc. Proceedings of the IEEE Conference Computer Vision Pattern Recognition, pp. 8183–8192 (2018)
13. Lai, W.S., Ding, J.J., Lin, Y.Y., Chuang, Y.Y.: Blur kernel estimation using normalized color-line prior. In: Proceedings of the IEEE Conference Computer Vision Pattern Recognition, pp. 64–72 (2015)
14. Lai, W.S., Huang, J.B., Hu, Z., Ahuja, N., Yang, M.H.: A comparative study for single image blind deblurring. In: Proceedings of the IEEE Conference Computer Vision Pattern Recognition, pp. 1701–1709 (2016)
15. Laine, S., Aila, T.: Temporal ensembling for semi-supervised learning. In: Proceedings of the International Conference Learning Representations (2017)
16. Levin, A., Weiss, Y., Durand, F., Freeman, W.T.: Understanding and evaluating blind deconvolution algorithms. In: Proceedings of the IEEE Conference Computer Vision Pattern Recognition, pp. 1964–1971 (2009)
17. Li, J., Wang, W., Nan, Y., Ji, H.: Self-supervised blind motion deblurring with deep expectation maximization. In: Proceedings of the IEEE/CVF Conference Computer Vision Pattern Recognition, pp. 13986–13996 (2023)
18. Liu, J., Yan, M., Zeng, T.: Surface-aware blind image deblurring. IEEE Trans. Pattern Anal. Mach. Intell. **43**(3), 1041–1055 (2019)
19. Luo, B., Cheng, Z., Xu, L., Zhang, G., Li, H.: Blind image deblurring via superpixel segmentation prior. IEEE Trans. Circuit Syst. Video Technol. **32**(3), 1467–1482 (2021)
20. Luo, F., Wu, X., Guo, Y.: AND: adversarial neural degradation for learning blind image super-resolution. In: Proceedings of the NeurIPS, pp. 21255–21267 (2023)
21. Michaeli, T., Irani, M.: Blind deblurring using internal patch recurrence. In: Proceedings of the European Conference Computer Vision, pp. 783–798 (2014)
22. Nah, S., Hyun Kim, T., Mu Lee, K.: Deep multi-scale convolutional neural network for dynamic scene deblurring. In: Proceedings of the IEEE/CVF Conference Computer Vision Pattern Recognition, pp. 3883–3891 (2017)
23. Pan, J., Su, Z.: Fast ℓ^0-regularized kernel estimation for robust motion deblurring. IEEE Sign. Process. Letters **20**(9), 841–844 (2013)
24. Pan, J., Sun, D., Pfister, H., Yang, M.H.: Deblurring images via dark channel prior. IEEE Trans. Pattern Anal. Mach. Intell. **40**(10), 2315–2328 (2017)
25. Perrone, D., Favaro, P.: Total variation blind deconvolution: the devil is in the details. In: Proceedings of the IEEE/CVF Conference Computer Vision Pattern Recognition, pp. 2909–2916 (2014)
26. Ren, D., Zhang, K., Wang, Q., Hu, Q., Zuo, W.: Neural blind deconvolution using deep priors. In: Proceedings of the IEEE/CVF Conference Computer Vision Pattern Recognition, pp. 3341–3350 (2020)
27. Ren, W., Tian, J., Tang, Y.: Blind deconvolution with nonlocal similarity and l_0 sparsity for noisy image. IEEE Sign. Process. Letters **23**(4), 439–443 (2016)
28. Ronneberger, O., Fischer, P., Brox, T.: U-Net: convolutional networks for biomedical image segmentation. In: Proceedings of the International Conference Medical Image Computing and Computer-Assisted Intervention, pp. 234–241 (2015)
29. Sun, J., Cao, W., Xu, Z., Ponce, J.: Learning a convolutional neural network for non-uniform motion blur removal. In: Proc. IEEE Conf. Comput. Vis. Pattern Recognit. pp. 769–777 (2015)
30. Sun, L., Cho, S., Wang, J., Hays, J.: Edge-based blur kernel estimation using patch priors. In: Proceedings of the IEEE International Conference Computer Photography, pp. 1–8 (2013)

31. Tao, X., Gao, H., Shen, X., Wang, J., Jia, J.: Scale-recurrent network for deep image deblurring. In: Proceedings of the IEEE Conference Computer Vision Pattern Recognition, pp. 8174–8182 (2018)
32. Tian, S., Zhang, S., Lin, B.: Blind image deblurring based on dual attention network and 2d blur kernel estimation. In: Proceedings of the IEEE International Conference Image Processing, pp. 1729–1733 (2021)
33. Ulyanov, D., Vedaldi, A., Lempitsky, V.: Deep image prior. In: Proceedings of the IEEE/CVF Conference Computer Vision Pattern Recognition, pp. 9446–9454 (2018)
34. Levin, A., Weiss, Y., Durand, F., Freeman, W.T.: Efficient marginal likelihood optimization in blind deconvolution. In: Proceedings of the IEEE/CVF Conference Computer Vision Pattern Recognition, pp. 2657–2664 (2011)
35. Wen, F., Ying, R., Liu, Y., Liu, P., Truong, T.K.: A simple local minimal intensity prior and an improved algorithm for blind image deblurring. IEEE Trans. Circuit Syst. Video Technol. **31**(8), 2923–2937 (2020)
36. Whyte, O., Sivic, J., Zisserman, A.: Deblurring shaken and partially saturated images. Int. J. Comput. Vis. **110**(2), 185–201 (2014)
37. Xu, L., Jia, J.: Two-phase kernel estimation for robust motion deblurring. In: Proceedings of the European Conference Computer Vision, pp. 157–170 (2010)
38. Xu, L., Zheng, S., Jia, J.: Unnatural l0 sparse representation for natural image deblurring. In: Proceedings of the IEEE/CVF Conference Computer Vision Pattern Recognition, pp. 1107–1114 (2013)
39. Zhang, J., et al.: Dynamic scene deblurring using spatially variant recurrent neural networks. In: Proceedings of the IEEE/CVF Conference Computer Vision Pattern Recognition, pp. 2521–2529 (2018)
40. Zhang, M., Fang, Y., Ni, G., Zeng, T.: Pixel screening based intermediate correction for blind deblurring. In: Proceedings of the IEEE/CVF Conference Computer Vision Pattern Recognition, pp. 5892–5900 (2022)
41. Zuo, W., Ren, D., Zhang, D., Gu, S., Zhang, L.: Learning iteration-wise generalized shrinkage-thresholding operators for blind deconvolution. IEEE Trans. Image Process. **25**(4), 1751–1764 (2016)

Complementary Dual-Branch Network for Space-Time Video Super-Resolution

Ming Tian, Tianyi Li, RongSheng Luo, Changxin Gao, and Nong Sang$^{(\boxtimes)}$

National Key Laboratory of Multispectral Information Intelligent Processing
Technology, School of Artificial Intelligence and Automation,
Huazhong University of Science and Technology, Wuhan 430074, China
{tianming,litianyi123,rongshengluo,cgao,nsang}@hust.edu.cn

Abstract. Space-time video super-resolution aims to simultaneously increase the space-time resolution of low-resolution and low frame-rate videos. Existing deep learning-based methods have made notable strides, predominantly achieving space-time video super-resolution through the relatively simple integration of modules for video super-resolution and video frame interpolation sub-tasks. However, these methods typically do not fully exploit the inherent relationships between the two sub-tasks. To address this limitation, we propose a Complementary Dual-Branch Network designed to better explore the interdependence of the two subtasks. Specifically, our dual-branch architecture facilitates mutual enhancement between video super-resolution and video frame interpolation subtasks within each branch and provides mutual guidance between the two branches. Additionally, we introduce a simple yet effective strategy for the rough estimation of optical flow, incorporating Flow-Guided Deformable Alignment into space-time video super-resolution to achieve precise motion estimation. In addition, we use an RNN-based Backward and Forward Recurrent module to ensure that all frames can utilize the information of the whole sequence. It is more efficient and memory saving compared to the currently popular bidirectional LSTM module. Experimental results on several datasets show that our method achieves superior accuracy and requires fewer parameters compared to state-of-the-art methods.

Keywords: Feature aggregation · Feature interpolation · Space-time video super-resolution

1 Introduction

With the growing popularity of advanced display technologies and the increasing demand for high-quality videos, space-time video super-resolution (STVSR) technology [23] has emerged. STVSR aims to leverage the spatial and temporal information in video sequences to generate high-resolution (HR), high-frame-rate videos (HFR) from given low-resolution (LR), low-frame-rate (LFR) videos.

Traditional STVSR methods have relied on manually designed regularization [6], prior knowledge such as the space-time directional smoothness prior

A. Antonacopoulos et al. (Eds.): ICPR 2024, LNCS 15332, pp. 182–195, 2025.
https://doi.org/10.1007/978-3-031-78125-4_13

[24], and assumptions like illumination consistency [20]. However, these methods often fall short in complex real-world scenarios. The advent of Convolutional Neural Networks (CNNs) have significantly advanced video enhancement tasks, including video super-resolution (VSR) [3,4,9,26,29,36], and video frame interpolation (VFI) [1,5,11,21,25]. Intuitively, executing VFI and VSR models separately on LR, LFR videos could achieve STVSR. However, these phased methods do not explore the interplay between temporal interpolation and spatial super-resolution. Moreover, these methods require separate model design and separate training for VFI and VSR, leading to redundancy, increased parameters, and reduced processing speed.

Recently, one-stage end-to-end models for STVSR [8,27,32,33,35] have gained popularity. These models extract features from the input LFR and LR frames, and then implement STVSR in the order of temporally interpolated frames and spatially aggregated super-resolution. However, similar to the phased methods, the interactions between VSR and VFI have not been fully investigated. These methods extract and aggregate features sequentially according to temporal and spatial dimensions, utilising only the help provided by VFI for VSR. Specifically, they leverage the additional temporal information generated by VFI to enhance the reconstruction of spatial details. These single-stage serial models also inevitably result in error accumulation. In addition, most alignment strategies of the current methods use either optical flow-based alignment or deformable convolution-based alignment. Explicit optical flow-based alignment is very dependent on the accuracy of motion estimation, and if the motion estimation is not accurate, the generated results are prone to have artifacts. Deformable convolution-based alignment is difficult to train in practice [3].

To overcome these issues, we introduce a Complementary Dual-Branch Network (CDBNet). Firstly, it implements mutual assistance between VSR and VFI within the two branches separately, using more temporal information generated by VFI for detail recovery in VSR and more spatial information generated by VSR for refinement in VFI. Secondly, linkages are established between the two branches, employing mutual guidance to mitigate the error accumulation inherent in the sequential models of each branch. Moreover, we propose a straightforward, efficient estimation strategy for estimating optical flow between missing and existing frames, and thus use the Flow-Guided Deformable Alignment (FDA) for frame alignment within our STVSR model. By using the coarse flow estimates between frames as the baseline part of the offset, the Deformable Convolution Network [37] (DCN) is required only to learn the residual of the offset, easing its training burden and promoting stable, quick convergence. In addition, for the feature aggregation and super-resolution, we use the excellent design of Backward and Forward Recurrent module to ensure that all frames utilize more information from the whole sequence. Connecting the RNN-based backward and forward module in series is more efficient compared to the module of parallel bidirectional ConvLSTM.

In this paper, we highlight our contributions as follows:

1) We introduce a Complementary Dual-Branch Network, which effectively leverages the synergy between the VSR and VFI and reduces error accumulation across the serial branches through inter-branch linkages.

2) We intorduce a practical estimation strategy for determining the optical flow of missing frames. This approach allows for the implementation of FDA in the STVSR task, leading to more accurate frame alignment.

3) Integrating the Backward and Forward Recurrent Module into CDBNet, our model significantly surpasses current state-of-the-art methods on different datasets, while maintaining a minimum parameter count.

2 Related Work

2.1 Video Super-Resolution

VSR aims to reconstruct high-resolution video from corresponding low-resolution video. For the VSR task, it is crucial to align features of neighboring frames with the reference frame and jointly extract their information to achieve super-resolution. Some methods [2,3,28,34] use optical flow for explicit alignment. However, with fast motion, optical flow alignment can be inaccurate, leading to artifacts. Therefore, TDAN [26] introduces deformable convolution to implicitly align inter-frame features, achieving impressive performance. EDVR [29] incorporates deformable convolution into a multi-scale pyramid module to further improve feature alignment. Moreover, combining the advantages of both, BasicVSR++ [4] proposes alignment based on both optical flow and deformable convolution. Recently, Liang et al. [14] realized alignment based on deformable attention [31].

2.2 Video Frame Interpolation

The goal of VFI is to synthesize an intermediate frame with two adjacent video frames, improving the temporal resolution of the video sequence. Some raditional methods based on path [18] and phase [19] struggle in complex scenarios. In recent years, deep learning-based methods have achieved significant success. Learning-based VFI methods can be categorized into flow-based methods and kernel-based methods. Flow-based methods [11,16,34] synthesize intermediate frames by estimating the optical flow between two frames and interpolating. SuperSlowMo [11] uses the U-net architecture to compute the optical flow between two frames. In addition, to address inaccuracies in optical flow estimation due to occlusion, DAIN [1] introduces a depth-aware module for occlusion detection. Kernel-based methods [5,21,25] use adaptive convolution to directly predict the kernels, which are then used to estimate intermediate frames. Niklaus et al. [21] use 1D kernels for adaptive convolution. To expand the receptive field, Cheng et al. [5] uses deformable and separable convolution.

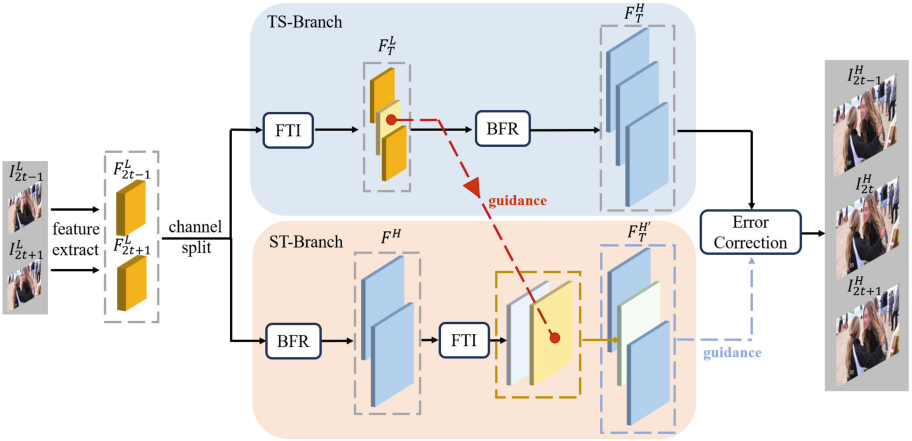


Fig. 1. Overview of the proposed CDBNet. For ease of drawing, we show only two input frames in the figure. It consists of two branches where the FTI and BFR modules are executed in different orders, and linkages are added between the two branches to guide each other.

2.3 Space-Time Video Super-Resolution

The goal of STVSR is to increase both spatial and temporal resolutions of LFR and LR videos. Recent deep learning-based work has made significant progress in this area. Haris et al. [8] propose an end-to-end network called STARnet, which achieves STVSR by additional optical flow branching to extract the association between temporal and spatial features, and jointly learning spatial and temporal content. xiang et al. [32] propose a ConvLSTM-based approach called Zooming Slow-Mo to align and interpolate intermediate features through deformable convolution. Xu et al. [33] introduce a temporal modulation block based on Zooming Slow-mo, thereby realizing time-controllable STVSR. LSTM architectures require significant memory to store intermediate states, making them less efficient than standard RNN architectures. Zhang et al. [35] propose an optical-flow-reuse-based bidirectional recurrence network, which balances the memory footprint and performance. Wang et al. [27] propose a deformable attention-based bidirectional network called STDAN. Geng et al. [7] propose a multiscale Transformer-based network called RSTT, which significantly reduces the number of parameters while maintaining similar performance to the above methods. However, all of these methods follow a single-branch serial structure that implements STVSR in the order of temporal frame interpolation and spatial super-resolution. The interaction between VSR and VFI subtasks has not been fully explored.

3 Method

3.1 Structure of the Complementary Dual-Branch Network

The architecture of our method is depicted in Fig. 1. First, we extract feature from the LR, LFR video sequence $\{I_{2t-1}^L\}_{t=1}^{N+1}$ using five residual blocks, resulting in the features $\{F_{2t-1}^L\}_{t=1}^{N+1}$. To reduce the number of parameters in the model, we divide the features F^L along the channel dimension and direct them into two branches. In TS-Branch, the Feature Temporal Interpolation (FTI) module conducts temporal alignment and interpolates frames within the features F^L, producing F_T^L. The refined temporal features F_T^L are then enhanced spatially through the Backward and Forward Recurrent (BFR) module to produce F_T^H. The super-resolution of details in the spatial dimension is helped by richer information in the temporal sequence. The process is formulated as:

$$\begin{aligned} F_{2t}^L &= FTI(F_{2t-1}^L, F_{2t+1}^L), \\ F_T^L &= \{F_t^L\}_{t=1}^{2N+1}, \\ F_T^H &= BFR(F_T^L). \end{aligned} \tag{1}$$

ST-Branch adopts the reverse process, which can help refine the temporal interpolation by using larger features in the spatial dimension.

Furthermore, we establish linkages between two branches to minimize error accumulation. The features interpolated by the FTI module of TS-Branch are upsampled and added to the features interpolated by the FTI module of ST-Branch to achieve guidance for interpolating the features of ST-Branch. The process is formulated as:

$$\begin{aligned} F^H &= BFR(F^L), \\ F_{2t}^H &= FTI(F_{2t-1}^H, F_{2t+1}^H) + \uparrow (F_{2t}^L), \\ F_T^{H'} &= \{F_t^H\}_{t=1}^{2N+1}, \end{aligned} \tag{2}$$

where ↑ denotes the upsampling operator. ST-Branch guides TS-Branch by concatenating its final features with those of TS-Branch. The combined features are processed through a convolution block for error correction, thus producing the final HR, HFR video sequence $\{I_t^H\}_{t=1}^{2N+1}$. The process is formulated as:

$$I^H = f_{3\times3}(F_T^H, F_T^{H'}), \tag{3}$$

where $f_{3\times3}$ denotes the convolution block for error correction.

3.2 Feature Alignment

Feature alignment is crucial in STVSR. For example, Zooming Slow-Mo [32] employs DCN for alignment, OFR-BRN [35] employs optical flow for alignment and the optical flow estimation of missing frames is achieved by IFnet [10]. However, both methods have limitations. Drawing inspiration from Basicvsr++

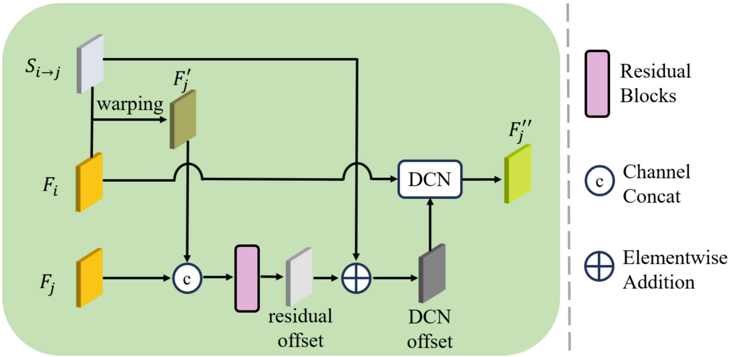


Fig. 2. An illustration of Flow-Guided Deformable Alignment (FDA).

[4] we use alignment based on FDA. In FDA, optical flow serves merely as a baseline for the offset, while the precise residual offset is determined through convolution. In STVSR, since half of the frames are missing, the optical flow of the missing frame and the existing frame cannot be directly obtained. It's necessary to first estimate the optical flow for both missing and existing frames. We employ pre-trained SPyNet [22] to calculate the optical flow of existing frames. We simplify the motion estimation between frames by assuming uniform linear motion, allowing us to approximate the optical flow between existing and missing frames by halving the optical flow betwee existing frames. The optical flow estimation can be formulated as:

$$S_{i\to i+1} \approx S_{i+1\to i+2} \approx \frac{1}{2} S_{i\to i+2}, \tag{4}$$

where $S_{i\to j}$ denotes the optical flow from i to j.

The FDA structure is shown in Fig. 2. We initially pre-align the feature F_i using the optical flow $S_{i\to j}$. Then calculate the residual part of offset by concatenating the feature F_j' with the feature F_j. We add up residual part of offset and the optical flow to get the DCN offset $o_{i\to j}$. Finally, applying the DCN to feature F_i to get the aligned feature F_j'', the process is formulated as:

$$\begin{aligned} &F_j' = warp(F_i, S_{i\to j}), \\ &o_{i\to j} = S_{i\to j} + ResBlock(Concat(F_j', F_j)), \\ &F_j'' = DCN(F_i, o_{i\to j}), \end{aligned} \tag{5}$$

where *warp* denotes the spatial warping operation.

3.3 Feature Temporal Interpolation

For the features F_{2t-1} and F_{2t+1} of the existing frames, we need to interpolate the feature F_{2t} of the missing frame. The interpolation process needs to draw on the forward and backward motion information between the existing and missing frames, and with the optical flow estimation strategy in the previous section, we can introduce the FDA to implicitly capture the forward and backward motion information, as shown in Fig. 3. After obtaining the two aligned features, we fuse them by channel cascading and a 1 × 1 convolution layer to obtain F_{2t}, the process is formulated as:

$$\begin{aligned} F_f &= FDA(F_{2t-1}, F_{2t+1}), \\ F_b &= FDA(F_{2t+1}, F_{2t-1}), \\ F_{2t} &= Fuse(F_f, F_b), \end{aligned} \tag{6}$$

where F_f denotes the feature containing forward motion information, F_b denotes the feature containing backward motion information and $Fuse$ denotes the 1×1 convolution layer.

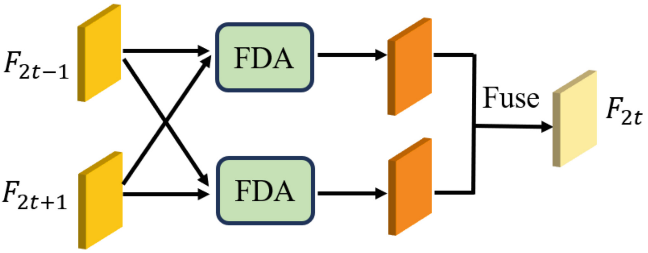


Fig. 3. An illustration of Feature Temporal Interpolation (FTI) module.

3.4 Backward and Forward Recurrent Module

Many previous works such as Zooming-Slow-Mo [32], TMNet [33] have highlighted the benefits of using a bidirectional ConvLSTM structure to capture information across entire video sequence. Despite its effectiveness, ConvLSTM requires significant memory due to the need to store multiple intermediate states, making it less efficient than vanilla RNN architecture. As depicted in Fig. 4, we utilize an RNN-like Backward and Forward Recurrent Module, where feature information from different frames is alternately propagated and extracted in the forward and backward branches. Both the forward and backward directions can utilise the information from the whole sequence. Compared to existing works, such as Zooming Slow-Mo and TMNet, which propagate features using simple parallel bidirectional architecture, this architecture can reduce the cumulative error of feature propagation in long sequences and improve feature expressiveness. f_i^2 is obtained by feeding f_{i-1}^2, which contains information about the forward sequence, and f_i^1, which contains information about the backward sequence, to the FDA for alignment extraction. The same process is applied to

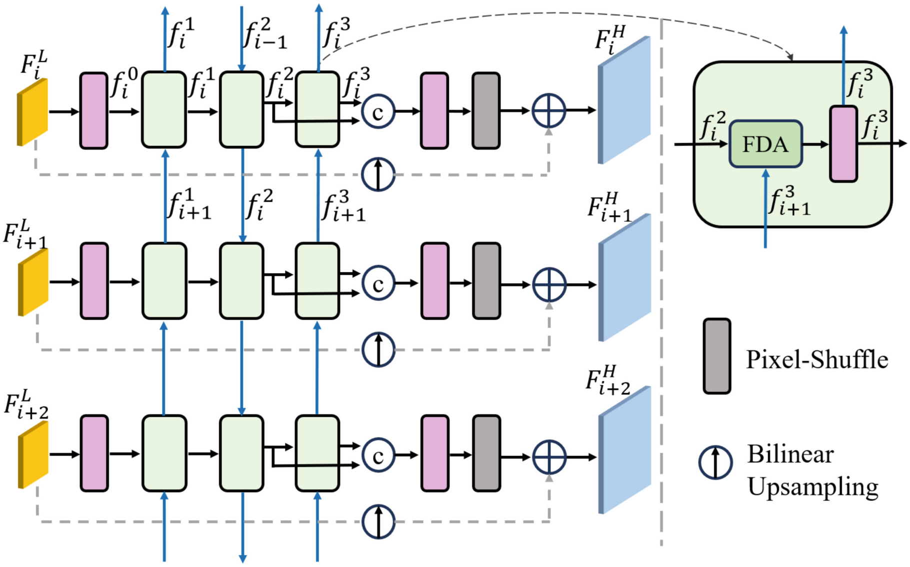


Fig. 4. An illustration of Backward and Forward Recurrent (BFR) module.

f_i^3. Thus, both f_i^2 and f_i^3 are generated utilizing information from the whole sequence. We concatenate the two and feed them into the residual block and Pixel-Shuffle to obtain the feature F_i^H with increased spatial size, the process is formulated as:

$$F_i^H = PS(ResBlock(Concat(f_i^2, f_i^3))) + \uparrow (F_i^L), \tag{7}$$

where PS denotes the Pixel-Shuffle operator.

4 Experiments

4.1 Implementation Details

For training, we utilized the Vimeo-90K dataset [34], which comprises 64,612 video sequences, each containing 7 consecutive frames. Our evaluation datasets include Vid4 [15] and the Vimeo-90K test set. To assess the performance across different motion scenarios, we divided the Vimeo-90K test set into three subsets based on motion speed: Fast, Medium and Slow, following the categorization in [32]. The LR frames were generated from the HR frames through bicubic interpolation by a factor of 4. For our experiments, we used the odd-indexed LR frames as inputs to predict continuous HR frames. The performance of various methods in STVSR was evaluated using Peak Signal-to-Noise Ratio (PSNR) and Structural Similarity Index [30] (SSIM) metrics, where higher values indicate closer resemblance to the ground truth.

We implemented our CDBNet in PyTorch and trained it on two NVIDIA GeForce RTX 4090 GPUs for 600,000 iterations. The Adam optimizer [12] was utilized in conjunction with the Charbonnier loss function [13] for optimization. The Charbonnier loss function can be formulated as:

$$L_{\text{rec}} = \sqrt{\|I_{\text{t}}^{GT} - I_{\text{t}}^{H}\|^2 + \epsilon^2}, \tag{8}$$

where I_{t}^{H} refers to the restoration outputs and I_{t}^{GT} denotes ground-truth HR video frames. ϵ is a constant value, and we empirically set it to $1e-3$. The initial learning rate was set to $4e-4$ and was gradually reduced to $1e-7$ using cosine annealing [17] every 150,000 iterations.

Table 1. Quantitative comparison of our method with other sota methods for STVSR. The best results are in bold and the second best results are with underline.

Method	Vid4		Vimeo-Fast		Vimeo-Medium		Vimeo-Slow		Speed	Parameters
	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	FPS	millions
SuperSloMo [11]+RBPN [9]	23.76	0.6362	34.73	0.9108	32.79	0.8930	30.48	0.8584	5.62	19.8+12.7
SepConv [21]+RCAN [36]	24.92	0.7236	34.97	0.9195	33.59	0.9125	32.13	0.8967	6.10	21.7+16.0
DAIN [1]+EDVR [29]	26.12	0.7836	35.81	0.9323	34.66	0.9281	33.11	0.9119	12.21	24.0+20.7
STARnet [8]	26.06	0.8046	36.19	0.9368	34.86	0.9356	33.10	0.9164	19.19	111.61
Zooming Slow-mo [32]	26.31	0.7973	36.81	0.9415	35.41	0.9361	33.36	0.9138	31.18	11.1
TMNet [33]	26.43	0.8016	37.04	0.9435	35.60	0.9380	33.51	0.9159	27.53	12.26
RSTT [7]	26.43	0.7994	36.80	0.9403	35.66	0.9381	33.50	0.9147	30.97	7.67
OFR-BRN [35]	26.72	0.8141	37.32	0.9465	**35.72**	0.9393	33.58	0.9167	40.12	11.77
CDBNet(Ours)	**26.83**	**0.8144**	**37.39**	**0.9512**	35.71	**0.9399**	**33.75**	**0.9194**	**49.31**	**5.05**

4.2 Comparisons with State-of-the-Art Methods

We benchmarked our method against both phased and one-stage end-to-end state-of-the-art (SOTA) methods. For phased methods, we implemented Sepconv [21], SupersloMo [11], DAIN [1] for VFI and RCAN [36], RBPN [9], EDVR [29] for VSR. One-stage end-to-end methods include STARnet [8], Zoom-Slow-Mo [32], TMNet [32], RSTT [7], OFR-BRN [35].

For a fair assessment of inference speed, all methods were evaluated on an NVIDIA GeForce RTX 4090 GPU. The quantitative outcomes are summarized in Table 1, demonstrating our method's superior performance in both PSNR and SSIM [30] across all datasets. Furthermore, our method is the fastest computationally (with the highest FPS) while minimizing the number of parameters. It can be seen that on the Vid4 and Vimeo datasets, our method outperforms the suboptimal method almost across the board in terms of PSNR and SSIM metrics, while running faster than it and with less than half the number of parameters. Figure 5 presents the qualitative results for the top-performing methods, demonstrating that our method excels in recovering the most accurate details. Our method generates better finger appearances and effectively improves aliasing due to poor alignment and error accumulation.

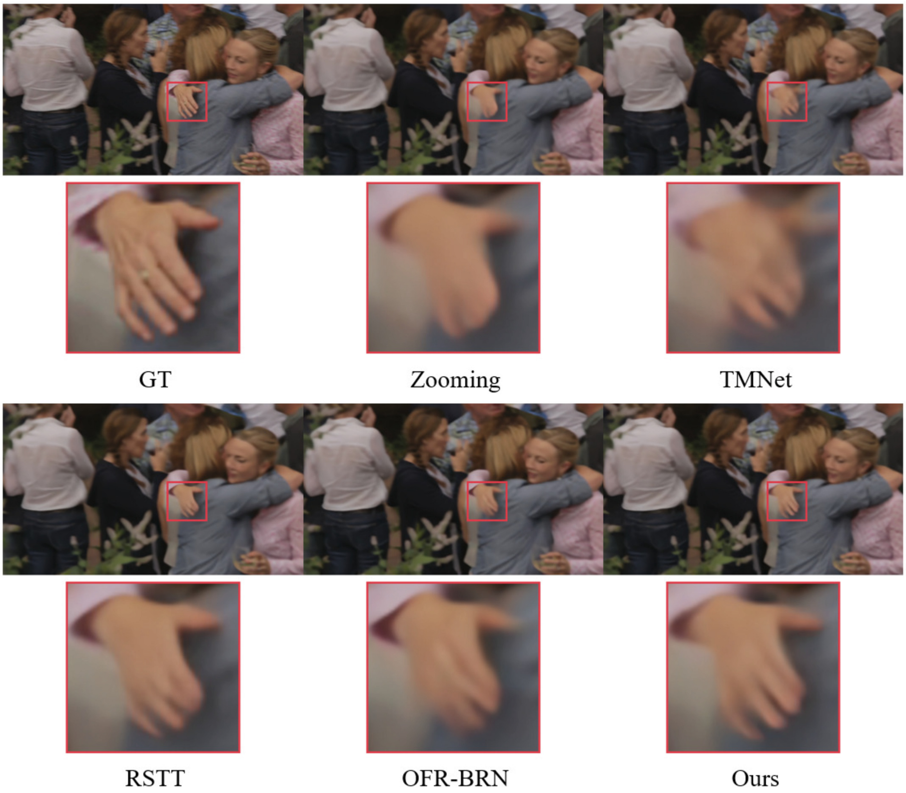


Fig. 5. Visual comparisons of different STVSR methods on Vimeo dataset. Parts of the areas are zoomed in and framed with red boxes to facilitate comparison. (Color figure online)

4.3 Ablation Study

We also tested the role of modules in our method. The results are shown in Table 2 and Fig. 6.

1) *The Structure of Dual-Branch Mutual guidance*: To evaluate the benefits of the dual-branch structure with mutual guidance, we modified the network structure to a single-branch structure in the order of temporal interpolation and spatial aggregation as well as a dual-branch structure without mutual guidance.

2) *The Flow-Guided Deformable Alignment*: To evaluate the benefits of incorporating FDA, we replaced the alignment module with a Pyramid, Cascading and Deformable (PCD) module for comparison.

The outcomes presented in Table 2 confirm that both the dual-branch architecture with mutual guidance and the FDA significantly enhance performance. The optimal results are achieved when integrating both elements. The results in Fig. 6 show that the structure generation using Single-Branch is visually poor, with the triangular distribution pattern on the backpack in the figure turning

Table 2. Ablation study on different modules.

Module	(a)	(b)	(c)	(d)
Dual-Branch		√	√	√
Mutual Guidance			√	√
FDA				√
Vid4 (PSNR)	26.04	26.25	26.59	26.83
Vimeo (PSNR)	35.05	35.22	35.44	35.57

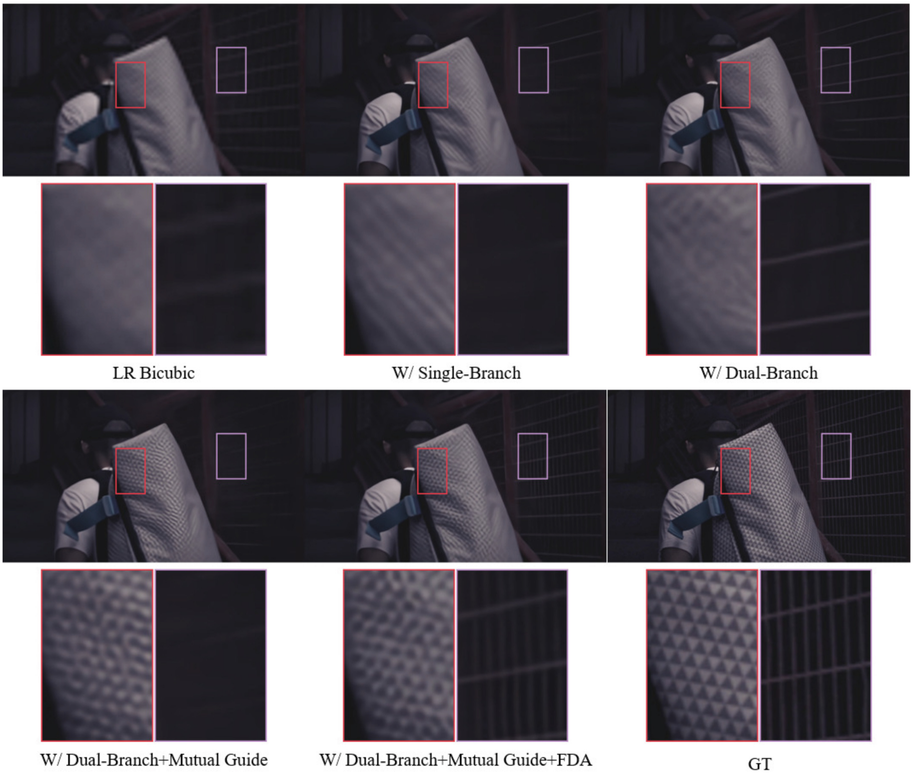


Fig. 6. Visual comparisons of ablation study on Vimeo dataset. Parts of the areas are zoomed in and framed with red and purple boxes to facilitate comparison. (Color figure online)

into a striped pattern. The result of structure generation using Dual-Branch can improve the error to some extent. After adding Mutual Guidance, the pattern pattern generation of the backpack was effectively improved. However, the generation results of the wooden windows have a mixing situation because of the inaccuracy of the inter-frame alignment. The best visual results were generated after using the FDA alignment on the structure of Dual-Branch and Mutual Guidance.

5 Conclusion

In this paper, we proposed a Complementary Dual-Branch Network (CDBNet) for STVSR. Our network has a dual-branch structure with links between branches that can fully exploit the relationship between the two subtasks of VSR and VFI and reduce error accumulation. Additionally, we have designed an efficient estimation strategy for the optical flow of missing frames, thus introducing FDA for more stable frame alignment. Experimental evaluations across various datasets have demonstrated that CDBNet outperforms current STVSR methods in terms of quantitative performance, visual quality, computational efficiency, and parameter count. In the future, we aim to extend our method to address a broader spectrum of real-world video degradation scenarios.

References

1. Bao, W., Lai, W.S., Ma, C., Zhang, X., Gao, Z., Yang, M.H.: Depth-aware video frame interpolation. In: Proceedings of the IEEE/CVF Conference Computer Vision Pattern Recognition, pp. 3703–3712 (2019)
2. Caballero, J., et al.: Real-time video super-resolution with spatio-temporal networks and motion compensation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4778–4787 (2017)
3. Chan, K.C., Wang, X., Yu, K., Dong, C., Loy, C.C.: BasicVSR: the search for essential components in video super-resolution and beyond. In: Proceedings of the IEEE/CVF Conference Computer Vision Pattern Recognition, pp. 4947–4956 (2021)
4. Chan, K.C., Zhou, S., Xu, X., Loy, C.C.: BasicVSR++: improving video super-resolution with enhanced propagation and alignment. In: Proceedings of the IEEE/CVF Conference Computer Vision Pattern Recognition, pp. 5972–5981 (2022)
5. Cheng, X., Chen, Z.: Multiple video frame interpolation via enhanced deformable separable convolution. IEEE Trans. Pattern Anal. Mach. Intell. **44**(10), 7029–7045 (2021)
6. Faramarzi, E., Rajan, D., Christensen, M.P.: Space-time super-resolution from multiple-videos. In: Proceedings of the International Conference Information Science Signal Processing Application, pp. 23–28 (2012)
7. Geng, Z., Liang, L., Ding, T., Zharkov, I.: RSTT: real-time spatial temporal transformer for space-time video super-resolution. In: Proceedings of the IEEE Conference Computer Vision Pattern Recognition, pp. 17441–17451 (2022)
8. Haris, M., Shakhnarovich, G., Ukita, N.: Space-time-aware multi-resolution video enhancement. In: Proceedings of the IEEE Conference Computer Vision Pattern Recognition, pp. 2859–2868 (2020)
9. Haris, M., Shakhnarovich, G., Ukita, N.: Recurrent back-projection network for video super-resolution. In: Proceedings of the IEEE/CVF Conference Computer Vision Pattern Recognition, pp. 3897–3906 (2019)

10. Huang, Z., Zhang, T., Heng, W., Shi, B., Zhou, S.: Real-time intermediate flow estimation for video frame interpolation. In: Proceedings of the European Conference Computer Vision, pp. 624–642. Springer (2022)
11. Jiang, H., Sun, D., Jampani, V., Yang, M.H., Learned-Miller, E., Kautz, J.: Super SloMo: high quality estimation of multiple intermediate frames for video interpolation. In: Proceedings of the IEEE Conference Computer Vision Pattern Recognition, pp. 9000–9008 (2018)
12. Kingma, D.P., Ba, J.: Adam: a method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
13. Lai, W.S., Huang, J.B., Ahuja, N., Yang, M.H.: Deep Laplacian pyramid networks for fast and accurate super-resolution. In: Proceedings of the IEEE Conference Computer Vision Pattern Recognition, pp. 624–632 (2017)
14. Liang, J., et al.: Recurrent video restoration transformer with guided deformable attention. Adv. Neural. Inf. Process. Syst. **35**, 378–393 (2022)
15. Liu, C., Sun, D.: On Bayesian adaptive video super resolution. IEEE Trans. Pattern Anal. Mach. Intell. **36**(2), 346–360 (2013)
16. Liu, Z., Yeh, R.A., Tang, X., Liu, Y., Agarwala, A.: Video frame synthesis using deep voxel flow. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 4463–4471 (2017)
17. Loshchilov, I., Hutter, F.: SGDR: stochastic gradient descent with warm restarts. arXiv preprint arXiv:1608.03983 (2016)
18. Mahajan, D., Huang, F.C., Matusik, W., Ramamoorthi, R., Belhumeur, P.: Moving gradients: a path-based method for plausible image interpolation. ACM Trans. Graph. (TOG) **28**(3), 1–11 (2009)
19. Meyer, S., Djelouah, A., McWilliams, B., Sorkine-Hornung, A., Gross, M., Schroers, C.: PhaseNet for video frame interpolation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 498–507 (2018)
20. Mudenagudi, U., Banerjee, S., Kalra, P.K.: Space-time super-resolution using graph-cut optimization. IEEE Trans. Pattern Anal. Mach. Intell. **33**(5), 995–1008 (2010)
21. Niklaus, S., Mai, L., Liu, F.: Video frame interpolation via adaptive separable convolution. In: Proceedings of the International Conference Computer Vision, pp. 261–270 (2017)
22. Ranjan, A., Black, M.J.: Optical flow estimation using a spatial pyramid network. In: Proceedings of the IEEE Conference Computer Vision Pattern Recognition, pp. 4161–4170 (2017)
23. Shechtman, E., Caspi, Y., Irani, M.: Increasing space-time resolution in video. In: Proceedings of the European Conference Computer Vision, pp. 753–768 (2002)
24. Shechtman, E., Caspi, Y., Irani, M.: Space-time super-resolution. IEEE Trans. Pattern Anal. Mach. Intell. **27**(4), 531–545 (2005)
25. Tao, X., Gao, H., Liao, R., Wang, J., Jia, J.: Detail-revealing deep video super-resolution. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 4472–4480 (2017)
26. Tian, Y., Zhang, Y., Fu, Y., Xu, C.: TDAN: temporally-deformable alignment network for video super-resolution. In: Proceedings of the IEEE/CVF Conference Computer Vision Pattern Recognition, pp. 3360–3369 (2020)
27. Wang, H., Xiang, X., Tian, Y., Yang, W., Liao, Q.: STDAN: deformable attention network for space-time video super-resolution. IEEE Trans. Neural Netw. Learn. Syst. (2023)

28. Wang, L., Guo, Y., Lin, Z., Deng, X., An, W.: Learning for video super-resolution through HR optical flow estimation. In: Computer Vision–ACCV 2018: 14th Asian Conference on Computer Vision, Perth, Australia, 2–6 December 2018, Revised Selected Papers, Part I 14, pp. 514–529. Springer (2019)
29. Wang, X., Chan, K.C., Yu, K., Dong, C., Change Loy, C.: EDVR: video restoration with enhanced deformable convolutional networks. In: Proceedings of the IEEE Conference Computer Vision Pattern Recognition Workshop, pp. 0–0 (2019)
30. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: from error visibility to structural similarity. IEEE Trans. Image Process. **13**(4), 600–612 (2004)
31. Xia, Z., Pan, X., Song, S., Li, L.E., Huang, G.: Vision transformer with deformable attention. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 4794–4803 (2022)
32. Xiang, X., Tian, Y., Zhang, Y., Fu, Y., Allebach, J.P., Xu, C.: Zooming Slow-Mo: fast and accurate one-stage space-time video super-resolution. In: Proceedings of the IEEE/CVF Conference Computer Vision Pattern Recognition, pp. 3370–3379 (2020)
33. Xu, G., Xu, J., Li, Z., Wang, L., Sun, X., Cheng, M.M.: Temporal modulation network for controllable space-time video super-resolution. In: Proceedings of the IEEE Conference Computer Vision Pattern Recognition, pp. 6388–6397 (2021)
34. Xue, T., Chen, B., Wu, J., Wei, D., Freeman, W.T.: Video enhancement with task-oriented flow. Int. J. Comput. Vis. **127**(8), 1106–1125 (2019)
35. Zhang, Y., Wang, H., Zhu, H., Chen, Z.: Optical flow reusing for high-efficiency space-time video super resolution. IEEE Trans. Circuit Syst. Video Technol. **33**(5), 2116–2128 (2023)
36. Zhang, Y., Li, K., Li, K., Wang, L., Zhong, B., Fu, Y.: Image super-resolution using very deep residual channel attention networks. In: Proc. Eur. Conf. Comput. Vis. pp. 286–301 (2018)
37. Zhu, X., Hu, H., Lin, S., Dai, J.: Deformable convnets V2: more deformable, better results. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 9308–9316 (2019)

Single-Image Driven 3D Viewpoint Training Data Augmentation for Effective Label Recognition

Yueh-Cheng Huang[1](✉), Hsin-Yi Chen[1], Cheng-Jui Hung[1], Jen-Hui Chuang[1], and Jenq-Neng Hwang[2]

[1] Department of Computer Science, National Yang Ming Chiao Tung University, Hsinchu, Taiwan
heaven830419@gmail.com

[2] Department of Electrical and Computer Engineering, University of Washington, Seattle, USA

Abstract. Confronting the critical challenge of insufficient training data in the field of complex image recognition, this paper introduces a novel 3D viewpoint transformation technique initially tailored for label recognition. This technique can be used not only for data augmentation by generating synthetic data from any perspective but also transform photos taken from any angle into a frontal view, thereby reducing the complexity of the recognition task. Given the extensive use of wine-related applications with over 20 million users and the continuous publication of wine label datasets, we decided to focus this study on wine labels. This method enhances deep learning model performance by generating visually realistic training samples from a single real-world label image, overcoming the challenges posed by the intricate combinations of text and logos. Unlike classical Generative Adversarial Network (GAN) methods, which fall short in synthesizing such intricate content combinations and require a large amount of training data to become effective, our proposed solution leverages time-tested computer vision and image processing strategies. By using just a single monocular wine label image, we can expand our training dataset, thereby broadening the range of training samples for deep learning applications. This innovative approach to data augmentation circumvents the constraints of limited training resources. We then utilize the augmented training images through the Vision Transformer (ViT) architecture, performing one-shot recognition of existing wine labels in the training classes or future newly collected wine labels unavailable in the training. Experimental results show a significant increase in recognition accuracy over conventional 2D data augmentation techniques, indicating the potential for broader application in various labeling scenarios.

Keywords: 3D viewpoint augmentation · Label recognition · Single-image training · Data synthesis · Frontalization

A. Antonacopoulos et al. (Eds.): ICPR 2024, LNCS 15332, pp. 196–211, 2025.
https://doi.org/10.1007/978-3-031-78125-4_14

1 Introduction

Label recognition system [1, 18] has become increasingly popular recently due to its practical usage. OCR (Optical Character Recognition) techniques are commonly used to extract text for wine label recognition, similar to offline handwriting methods [14]. Deep learning-based scene text detection methods [6, 7, 21–23] further help identify text regions [28]. However, challenges arise from defaced text [17], language variations, mixed fonts, and intertwined text and graphics [15], complicating the recognition [25]. As a result, recently research works [18–20, 34] are focus on using image-based method rather than text-based method such as OCR, due to the image-based recognition system can build end-to-end model only by images. However, learning-based methods always face the challenge of having insufficient training data, which remains to be a critical issue for such models to perform satisfactorily. Regardless of the improvements in model design and training techniques, using insufficient and unrepresentative data for training can result in inadequate performance of generalization [8, 24]. Also, obtaining enough large and diverse training data that are representative of the target dataset remains a challenging task for many practical applications [11, 16, 27]. Moreover, when applying deep learning to real-world tasks, it is often encountered that the training data are very different from the test data.

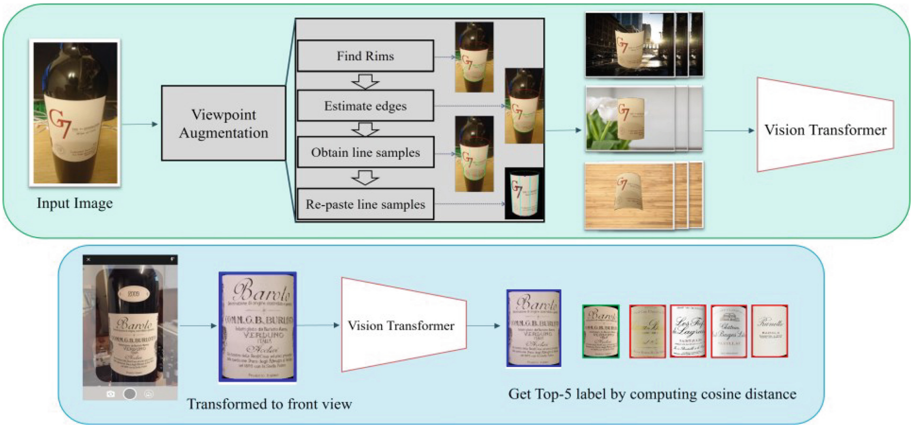


Fig. 1. Pipeline of wine label recognition with limited data. The upper green section illustrates the process of generating synthetic data using the 3D viewpoint transformation technique, which serves as training data for the model. The lower blue section represents the inference stage, where the 3D viewpoint transformation is utilized to simplify the testing of the model. This allows the model to more easily identify the corresponding labels based on cosine similarity.

To address the challenges in label recognition, this paper justifies focusing on wine labels due to their significant market presence, as evidenced by high sales in the beverage sector and over twenty million downloads of wine recognition apps [1]. We propose an innovative 3D viewpoint augmentation pipeline that generates a diverse and realistic training dataset from a single label image. This method effectively trains a deep learning model for label recognition, as illustrated in Fig. 1. Moreover, to further enhance the robustness of our label recognition system, particularly during the inference stage, we also implement a 3D viewpoint transformation as well during inference step. Each photo captured is systematically transformed into a consistent frontal view and same-distance perspective prior to the template matching process. This standardization is crucial as it mitigates the variable factors associated with user-taken photographs, such as differing angles and distances, which can significantly impact the recognition accuracy.

The ViT model with an MLP head is typically trained to classify input images into predefined categories. However, in the case of recognition, the market is continuously introducing new varieties, often with subtle variations, such as different vintages [4]. A significant challenge with direct classification is the model's adaptability to these new labels without the need for retraining. Therefore, for wine label recognition, we employ metric learning based one-shot recognition, which involves comparing the similarity of feature vectors (embeddings) from the frontal view test data with those of the original training data embeddings.

By combining 3D viewpoint training data augmentation for metric learning of embedding features, we have developed an efficient and precise wine label recognition system. This approach not only addresses the shortcomings of traditional methods in recognizing new varieties of wine labels but also demonstrates the tremendous potential of applying deep learning techniques in rapidly changing product categories. Contributions of the paper include:

(i) Our label recognition pipeline can be effectively trained with very limited training data, requiring as few as only one sample.
(ii) Our proposed 3D viewpoint transformation method can transform any taken photo into a consistent front and same-distance view. Not only does it save the labor of manually capturing various angles, but it also enhances accuracy.
(iii) Our proposed 3D viewpoint data augmentation for metric learning can improve the Top-1 accuracy significantly, i.e., more than 13.8%, over the standard 2D data augmentation based deep learning model.

2 Related Work

2.1 Data Augmentation

The challenge of having insufficient training data remains a critical issue for such a model to perform satisfactorily. Regardless of the improvements in model design and training techniques, using insufficient and unrepresentative data for training can result in inadequate performance of generalization [8, 24]. Also, obtaining enough large and diverse training data that are representative of the target dataset remains a challenging task for many practical applications [11, 16, 27]. Moreover, when applying deep learning to real-world tasks, it is often encountered that the training data are very different from the test data. Many works [10, 12] have shown that appropriate data augmentation techniques can help address this issue by generating additional training samples from the existing ones. Specifically, as described in [32], image data augmentation approaches can be roughly classified into two categories, which are: (i) based on basic image manipulations or (ii) based on deep learning. For (i), image augmentations can be carried out by geometric transformation (flipping, cropping, rotation, translation, shearing), noise injection, random erasing, color space transformation, and image mixing. Such manipulations try to preserve the main features of existing images from the training dataset, while adding potential variations for better generalization in the test dataset. Similarly, deep-learning-based augmentations of (ii), such as adversarial training, style transfer, and generative adversarial networks (GANs), exploit CNN-based network(s) to achieve a variety of image styles, e.g., changing lighting directions and intensities, or generate realistic images through learned image features from the training dataset [2]. While existing data augmentation methods enhance test accuracy in various learning-based tasks, they often fall short in more specialized applications [3]. Specifically, standard image manipulations fail to produce realistic images for certain tasks [30], while deep-learning-based augmentations necessitate additional, specific training images [5].

These limitations are especially pronounced in the task of wine label recognition, where conventional methods cannot adequately simulate the realistic perspective of labels on cylindrical wine bottles, while advanced methods require diverse images from multiple viewpoints for effective training. Therefore, our 3D augmentation method aims to address the inefficiencies and inaccuracies of traditional approaches, as well as the time-consuming and labor-intensive of learn-methods that require extensive training data.

2.2 Frontalization

Frontalization techniques, originally developed for facial recognition [32, 33], involve transforming images of faces from any viewpoint into a standardized, frontal position. This transformation significantly enhances the performance of face recognition systems by effectively handling variations in lighting, facial expressions, and occlusions. Its effectiveness in real-world scenarios, where such inconsistencies are prevalent, underscores the robustness and practicality of this method for improving face recognition tasks. To achieve this, the process begins with 2D alignment, which identifies six key fiducial points on the face, such as the centers of the eyes, the tip of the nose, and mouth locations. These points are utilized to scale, rotate, and translate the image through a series of transformations [31]. However, to overcome the limitations of 2D alignment, especially with out-of-plane rotations which are crucial for accurate frontalization, the process extends into 3D alignment. The method proposed in [34] uses a generic 3D shape model and a 3D affine camera to project the 2D aligned image onto a 3D plane. This sophisticated approach not only enhances alignment by incorporating additional fiducial points into the 3D model but also ensures an accurate correspondence between detected and reference points. The affine transformation is further optimized through a loss function that takes into account the covariances of fiducial point locations, ensuring the frontalization is precise and reliable.

Adapted for wine labels, our method involves reorienting images of labels captured from various angles to a consistent, front-facing view. By standardizing the orientation of wine labels at the inference stage, this technique ensures that all labels are uniformly aligned and presented, thus enhancing the model's ability to recognize and process them accurately.

3 Proposed Method

3.1 3D Viewpoint Augmentation from a Single Image

In this section, the proposed 3D viewpoint data augmentation scheme for a single wine label image is presented. This scheme needs to estimate, to some extent, the corresponding pose of the cylindrical bottle, and generate perspective realistic texts and patterns of the wine label in the augmented image. As shown in the green upper section of Fig. 1, the 3D viewpoint augmentation consists of three critical steps:

2D Description of 3D Surface: The process begins with converting the 3D cylindrical surface of the wine label into a 2D representation. This involves identifying both the upper and lower elliptical rims (latitudinal edges) and the two straight longitudinal edges.

Line Sample Extraction: Next, we use the vanishing point from the above longitudinal edges to extract 2D line samples along the label's longitudinal direction.

Perspective Mapping: The final step involves mapping these line samples onto an image of a cylindrical surface with a different pose. This mapping, which uses a view-invariant cross-ratio technique, ensures the correct perspective of the wine label on each line sample.

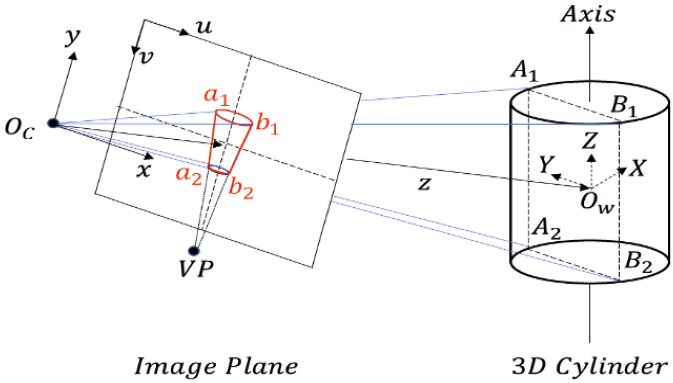


Fig. 2. Projective geometry of a cylinder

3.2 Projective Geometry of a Cylinder

According to projective geometry, images of the (circular) top and bottom plates of a (3D) cylinder, as illustrated in Fig. 2, will have elliptical shapes. As for the two edges of the cylinder in the image, i.e., $\overline{a_1a_2}$ and $\overline{b_1b_2}$, they correspond to the intersections of the image plane and planes $O_CA_1A_2$ and $O_CB_1B_2$, respectively, with both planes tangent to the 3D cylinder and passing through the camera center (O_C). Moreover, the intersection of $\overline{a_1a_2}$ and $\overline{b_1b_2}$ corresponds to the vanishing point (VP) of all 3D lines parallel to the axis of the cylinder.

3.3 Obtaining the Image Region of a Rectangular Wine Label Pasted on a Bottle

In this section, a 2D geometric description of the image of a rectangular wine label pasted on a 3D cylindrical surface (of a wine bottle) is provided. Such description will be used in the next section to obtain 1D (longitudinal) line samples of the wine label region.

3.3.1 Deriving Upper and Lower (Elliptical) Rims of the Wine Label Region

For the geometry shown in Fig. 2, elliptical expressions of the upper/lower rims of a roughly vertically oriented wine label region, e.g., for the image shown in Fig. 3(a), can be obtained with the following procedure:

1. Convert the color image to a gray-level image (Fig. 3(b))
2. Identify edge pixels with a large image gradient in the vertical direction (Fig. 3(c))
3. Identify image blocks, i.e., edge blocks, with enough (relative to block dimension) edge pixels (Fig. 3(d))
4. Label positive/negative (red/green) edge block according to the gradient direction of most edge pixels (Fig. 3(e))
5. Establish the longest chains of positive, and negative, edge blocks (Fig. 3(f))
6. Obtain (thinned) rim pixels by performing non-maximum suppression in the vertical direction (Fig. 3(g))
7. Obtain elliptical expressions of the upper and lower rims via curve fitting (Fig. 3(h))

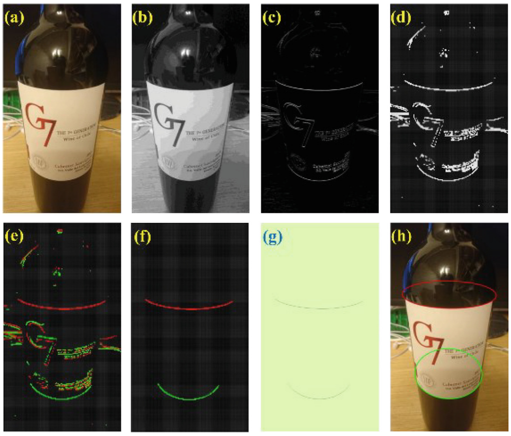


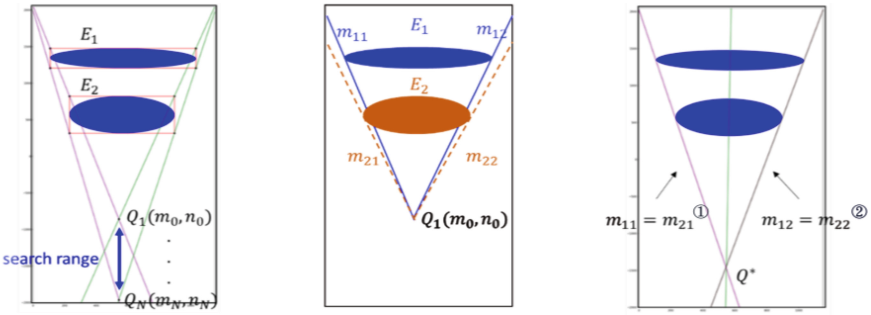
Fig. 3. Obtaining the elliptical expressions of the upper and lower rims of a wine label region (see text).

While the elliptical expressions in Step 7 can be obtained, for the rim pixels identified in Step 6, with an OpenCV function, default parameters need to be selected for some of the above processes, including minimum gradient (80, Step 2), block size (1/80 of image width, Step 3), minimum edge pixels (60% of block width, Step 3), and maximum gap in the chain (2 blocks, Step 5). Like a typical image processing procedure, different parameters may need to be determined, possibly manually, for some extreme imaging conditions. Nonetheless, such effort is worthwhile as unlimited synthetic, and perspective realistic, images can be obtained from such data augmentation, which will ultimately benefit the subsequent task of wine label recognition.

3.3.2 Obtaining the Left and Right Edges of the Wine Label

As described in Sect. 3.2, left and right (longitudinal) edges of the wine label in an image correspond two common external tangents of the two ellipses obtained in Sect. 3.3.1. The procedure of finding these edges for two ellipses, e.g., E_1 and E_2 depicted in Fig. 4(a), can be summarized as follows:

1. Identify search range for their intersection (VP) using bounding boxes of E_1 and E_2 (Q_1 to Q_N in Fig. 4(a)).
2. Obtain initial four tangents to E_1 and E_2 with Q_1 (m_{11}, m_{12}, m_{21}, and m_{22} in Fig. 4(b)).
3. Obtain the two common external tangents via binary search ($m_{11} = m_{21}$ and $m_{12} = m_{22}$ in Fig. 4(c)).



(a) (b) (c)

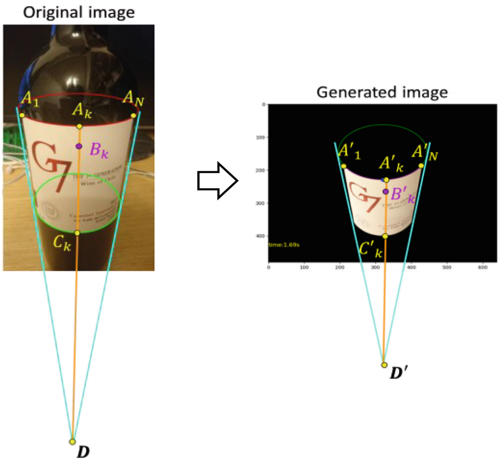
Fig. 4. Obtaining left and right edges of a wine label region.

For Step 2, two tangents of an ellipse from an arbitrary point outside of the ellipse can be obtained analytically, which are omitted here for brevity. As for Step 3, it is not hard to see that the slopes of all tangent lines will change monotonically with respect to the location of their intersection along a line and will not be the same except for a common external tangent; therefore, binary search can be employed to solve the problem efficiently.

3.4 Obtaining 2D (Longitudinal) Line Samples

To synthesize the image for a novel view of a wine label, point samples of the label need to be obtained from a real image captured in advance. To facilitate the geometrically natural synthesis process presented in the next subsection, these samples will first be obtained along the longitudinal direction of the wine label, i.e., is parallel to the axial direction of the wine bottle. As parallel lines in the 3D space will intersect at a VP (Sect. 3.2) in an image, the following sampling scheme is adopted (see Fig. 5(a)).

1. Identify the two common external tangents' intersection as the VP (D in Fig. 5(a)).
2. Identify the wider rim, and its leftmost/rightmost rim pixels (A_1 and A_N in Fig. 5(a)).
3. Obtain line samples by connecting rim pixels between A_1 and A_N toward D, e.g., $\overline{A_k C_k}$ is identified as the k-th line sample with C_k belonging to the smaller rim.



(a) (b)

Fig. 5. (a) Obtaining (longitudinal) line samples of a wine label region, and (b) re-projecting them onto a virtual (invisible) wine bottle with an arbitrary pose (see text).

3.5 Synthesizing Wine Label Images for Perspective Realistic Data Augmentation

Once foregoing line samples are obtained, they can be pasted onto the image of a cylindrical surface obtained from a novel view of the wine bottle, possibly via perspective projection of a graphic model. In this paper, such process is performed by pasting these line samples one at a time, with nonlinear mapping of pixel locations (based on the view-invariant cross-ratio between these locations) along each line according to the geometry of perspective projection, as shown in Fig. 5(b), so as to achieve perspective realistic appearance of the resultant synthetic image. In particular, the above re-projection process can be summarized as follows.

1. Identify the two common external tangents' intersection as the VP (D' in Fig. 5(b)).
2. Identify the wider rim, and its leftmost/rightmost rim pixels (A'_1 & A'_N in Fig. 5(b)).
3. Re-project image pixels of each line sample, e.g., $\overline{A_k C_k}$ in Fig. 5(a), to the corresponding line segment connecting the two new rims, e.g., the k-th line segment $\overline{A'_k C'_k}$ in Fig. 5(b), using the cross-ratios.

While Steps 1 and 2 are like their counterparts in Sect. 3.4, locations of A'_1 and A'_N in Fig. 5(b), and thus the number of line samples need to be synthesized in Step3, need to be determined. As we have just a single image of wine label, i.e., the image shown in Fig. 5a), without having other camera/environmental information, these two locations are approximately estimated with respect to the width of the larger rim.[1] As for the re-projection performed in Step 3, since the location of three points are already determined along $\overline{A'_k C'_k}$, any image pixel of $\overline{A'_k C'_k}$ can be determined by solving the following equation of view-invariant cross-ratio, with $B\prime_k$ being the only unknown.

$$\frac{A'_k C'_k \cdot B'_k D'}{A'_k D' \cdot B'_k C'_k} = \frac{A_k C_k \cdot \boldsymbol{B_k} D}{A_k D \cdot \boldsymbol{B_k} C_k} \tag{1}$$

As the geometry of perspective projection is approximately satisfied in the foregoing process of re-projection, numerous visually realistic images of wine label can be generated from a real wine label image. Figure 6 shows some synthetic images thus obtained from a single wine label image; wherein only one-dimensional rotation/translation of the virtual wine bottle is considered in each image so that the variation of its pose can be observed more easily. Note that the foregoing results are based on a virtual camera system which is established to mimic the imaging process of a typical cell phone camera. In particular, a virtual wine bottle with diameter equal to 76 mm is placed about 150 mm in front of the camera which has a focal length of about 6.8 mm.

3.6 Embedding Features from Metric Learning of a ViT

Recently, ViT [13] has achieved superior results in computer vision compared to traditional CNN-based approaches. Its ability to exploit global contextual information, coupled with its strong representation learning capabilities, makes it particularly suitable for wine label classification tasks.

3.6.1 Training and Testing Procedure

During training stage of ViT, we employ 3D viewpoint augmentation to augment our training data, obtaining 2D images of wine labels observed from different perspectives. Second, we train the model using the augmented images through metric learning to obtain the discriminative embeddings of wine labels. The objective of the model is to minimize the cosine distance between embeddings of wine labels of the same class, while simultaneously increasing such distance between embeddings of different classes. For the testing stage, our model primarily relies on the embedding feature representation of a single 3D viewpoint generated frontal view image, even though we have expanded the dataset through data augmentation. Therefore, during similarity calculations, we do not compare with all the embeddings or the average embedding of all augmented training data. Instead, we focus on comparing with the embedding of the single frontal view 3D viewpoint generated image. Moreover, considering the importance of the original data's

[1] Although further investigation is still needed for such issue, images synthesized with these simple estimations seem to work satisfactorily, as will be demonstrated in the experimental results.

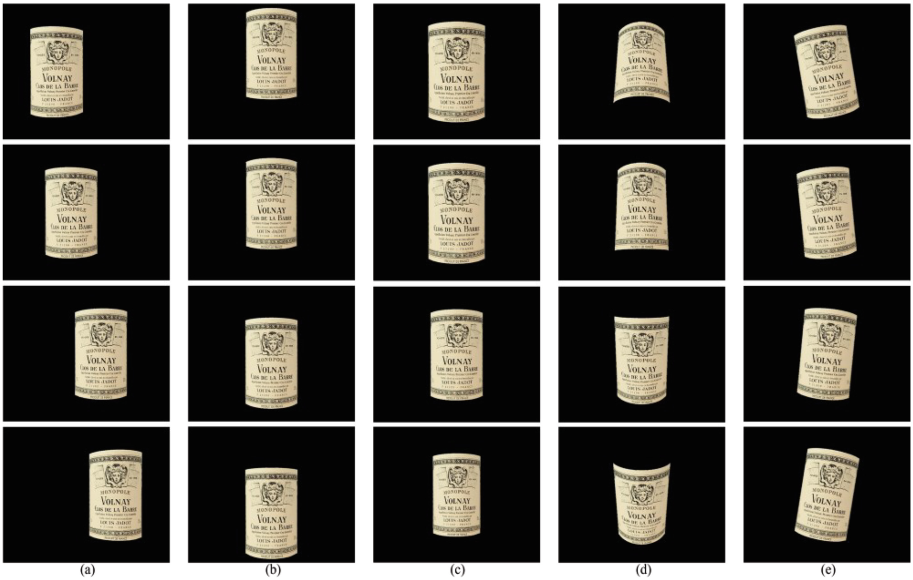


Fig. 6. Visually realistic (640 × 480) images synthesized upon a virtual (invisible) wine bottle which is translated between (a) $x = -40$ mm and $x = 40$ mm, (b) $y = -20$ mm and $y = 20$ mm, and (c) $z = 230$ mm and $z = 270$ mm and rotated between (d) − 30° and 30° w.r.t. the x-axis, and (e) − 10° and 10° w.r.t. the z-axis.

quality on model performance, we use the method mentioned above to ensure test data are compared with the embedding of a 3D viewpoint generated frontal view image. This method is particularly crucial for handling wine labels with subtle variations, as it allows the model to accurately identify new or slightly altered labels based on a reliable and consistent reference point.

3.6.2 ViT Dino and Loss Function

In our study, we have adopted the ViT architecture [13] which has been further advanced in the context of self-supervised learning within the DINO framework by Caron et al. [9]. Our approach is in line with the implementation utilized in DeiT [29], known for its effectiveness across a range of image processing tasks. For loss function, we use batch all triplet loss strategy proposed by [14], which is a variation of the conventional triplet loss [26].

With triplet loss, given an anchor sample x_a, the projection distance D of a positive sample x_p belonging to the same class x_a should be closer to the anchor's projection than that of a negative sample belonging to a different class x_n, by at least a margin m. On the other hand, the batch all triplet loss aims to enhance the efficiency and effectiveness of training deep metric learning models. The batch all triplet loss, denoted as $\mathcal{L}_{\text{BA}}$, involves forming batches by randomly selecting P classes (wine identities) and randomly sampling K images from each class (wine). Then, it computes the triplet loss for all possible combinations of triplets, given by:

$$\mathcal{L}_{\mathrm{BA}}(\theta; X) = \overbrace{\sum_{i=1}^{P}\sum_{a=1}^{K}}^{all\,anchors} \overbrace{\sum_{\substack{p=1\\p\neq a}}^{K}}^{all\,pos.} \overbrace{\sum_{\substack{j=1\\j\neq i}}^{P}\sum_{n=1}^{K}}^{all\,negatives} \left[m + d_{j,a,n}^{i,a,p}\right],$$

where

$$d_{j,a,n}^{i,a,p} = D\left(f\left(x_a^i\right), f\left(x_p^i\right)\right) - D\left(f\left(x_a^i\right), f\left(x_n^j\right)\right).$$

3.6.3 Inference Stage

During the training phase, due to our 3D viewpoint transformation technique, we generate a large number of augmented copies. This is crucial as it ensures that we have ample training data with slight deviations, allowing the deep learning model to learn from a broader range of perspectives and nuances. This enriched training phase is designed to robustly prepare the model for diverse real-world scenarios.

On the other hand, during the testing phase, we use the same 3D viewpoint transformation technique to simplify the inference stage by standardizing all test data to a frontal view with the orientation set to zero degrees ($x, y, z = 0$). This standardization is instrumental in reducing unnecessary noise and variability in the inference stage, thus enabling more consistent and accurate model predictions. By aligning all test images to a uniform orientation, we mitigate the impact of angle variations and ensure that the model's performance is evaluated based on its ability to recognize and process the essential features of the input data without the confounding factor of orientation differences. This approach helps in achieving higher precision and reliability in the model's output during real-world applications.

4 Experiments

In this study, we utilize WineSensed [34], an extensive multimodal wine dataset, to explore the relationships between visual perception, language, and flavor. This dataset comprises approximately 897,000 wine label images sourced from the Vivino platform. For our wine label recognition experiments, we selected classes from WineSensed that contain more than two images. We used one image per class with our 3D augmentation for training and the remaining images for testing, totaling 32,217 classes.

4.1 Improvement Achieved Through Viewpoint Augmentation

Firstly, we evaluated the performance of various deep learning models, including several ViT architectures, on the wine dataset using conventional 2D data augmentation, advanced 2D techniques such as CutOut, Mixup, and CutMix, as well as our 3D viewpoint augmentation. Table 1 provides a comprehensive evaluation of the wine label recognition accuracy. It is readily observable that our 3D viewpoint augmentation, which produces visually more realistic images, outperforms not only the conventional 2D data augmentation but also the more advanced 2D techniques by a significant margin. Specifically, there is up to a 13.8% improvement in the Top-1 accuracy results for the ViT-S/16 model when using 3D viewpoint augmentation, indicating that more relevant image embeddings can be generated with our 3D scheme compared to both basic and advanced 2D data augmentations.

Table 1. Performance comparison of 2D traditional augmentation method and our 3D viewpoint augmentation method across different models.

Condition	2D Augmentation		Advanced 2D Augmentation		3D Augmentation	
	Top-1 Acc.	Top-5 Acc.	Top-1 Acc.	Top-5 Acc.	Top-1 Acc.	Top-5 Acc.
VIT-S/16	70.8%	80.3%	74.3%	83.4%	**84.6%**	**90.5%**
VIT-S/8	67.3%	77.8%	71.5%	80.1%	82.3%	89.2%
VIT-B/16	70.4%	80.7%	73.5%	82.7%	84.3%	90.4%

When comparing the performance of ViT with our proposed data augmentation and metric learning techniques, we observe a significant improvement facilitated by our 3D viewpoint augmentation for ViT. Specifically, taking ViT-S/16 as an example, it achieves remarkably high performance. This can be attributed to the higher discriminative power of the embeddings generated by ViT, as depicted in Fig. 7. The heatmap illustrates that ViT's attention mechanism is not concentrated on a single point, but rather distributed across multiple regions of the image. This characteristic enables ViT to capture subtle differences in wine labels, including variations in textures, fonts, and design elements. Through metric learning, this distributed attention pattern further enhances the model's classification capabilities.

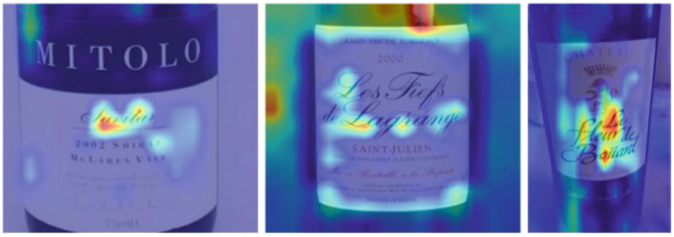


Fig. 7. Heatmaps from ViT-S/16 showing embeddings for various input images.

4.2 Enhancing Wine Label Recognition with Background Replacement

3D viewpoint augmentation involves augmenting the wine bottle data by introducing different poses of the bottles. However, it only generates foreground wine labels, while the background remains unspecified (black). Having a purely black background not only fails to fully exploit the data synthesis characteristics but also increases the risk of model overfitting. Therefore, we replace the black background with randomly sourced background images from the internet, as shown in Fig. 8. It is shown in Table 2, after such replacement, the accuracy of recognition may increase by 4.1%. The reason behind such an improvement is that complex backgrounds introduce additional variations and noises into the data. This challenges the model to discern relevant foreground features and learn to focus on the important aspects of the input. By learning to ignore or adapt to complex backgrounds, the model becomes more robust in distinguishing signal from noises.

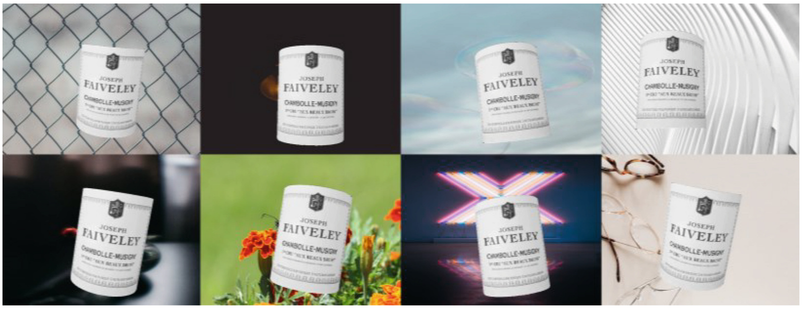


Fig. 8. Image synthesis examples with black background regions replaced by random background images.

Table 2. Accuracy of ViT-S/16 in wine label recognition using 3D viewpoint augmentation, and for different backgrounds.

ViT-S/16 + 3D Aug	Condition	Top-1 Acc.	Top-5 Acc.
	w/o Background Replacement	80.4%	87.3%
	w/ Background Replacement	**84.6%**	**90.5%**

4.3 Comparative Analysis of Perspective Transform

2D perspective transform is a commonly used data augmentation technique that allows images to be transformed from one perspective to another. Similar to our approach, it can alter the position, angle, and size of objects in an image to simulate the viewpoint of an observer to some extent. Nonetheless, due to the special curved surface of wine labels cropped from wine bottles, our 3D viewpoint augmentation can create more realistic augmentation of wine label images than those created by 2D perspective transform.

To show the performance difference of 2D perspective transform compared with our method, we conduct experiments on the ViT-S/16 model for different degrees of perspective transforms. As shown in Table 3, where all our 3D viewpoint augmentation results outperform the 2D augmentation with different settings of perspective transform. Interestingly, while adding more 2D perspective transformations will indeed improve the accuracy for 2D augmentation methods in the ViT-S/16 model, it may actually have negative impacts on the performance of our 3D augmentation results. For example, our best results for the Top-1 accuracy are achieved by skipping the 2D perspective transformation completely.

Table 3. Accuracy of wine label recognition using ViT-S/16 and different perspective transformation schemes.

Condition	2D Augmentation		3D Augmentation	
	Top-1	Top-5	Top-1	Top-5
ViT-S/16 + Little Perspective Aug	70.8%	80.3%	84.6%	**90.5%**
ViT-S/16 + Big Perspective Aug	76.6%	82.7%	82.7%	87.9%
ViT-S/16 + No Perspective Aug	–	–	**85.5%**	90.3%

4.4 Frontal Test Data for Enhanced Accuracy

In pursuit of higher accuracy in wine label recognition, we also employ the 3D viewpoint transformation method described in Sect. 3 for preprocessing the test data, converting wine labels into a frontal view 3D viewpoint generated image. This step is particularly crucial for handling real-world wine label images, which are often captured by users under less-than-ideal conditions, resulting in images that may be skewed or contain a high level of noise. By transforming these images into a standardized frontal view before testing, we can significantly enhance the accuracy of our wine label recognition model. As Table 4 shown, there is an improvement of 3% in Top-1 accuracy and 4.5% in Top-5 accuracy, indicating that the ViT model can more precisely recognize the correct label with this step. This ensures that even real-world, imperfect images are accurately recognized by the deep learning model.

Table 4. Accuracy of ViT-S/16 in wine label recognition task using 3D view-point transformation for test data.

ViT-S/16 + 3D Aug	Condition	Top-1 Acc.	Top-5 Acc.
	w/o transformed frontal-view test data	84.6%	90.5%
	w/ transformed frontal-view test data	87.2%	94.9%

5 Conclusion

Data augmentation is a way to extend training data so that deep learning models can achieve good results in situations where such data are limited, of poor quality, or even absent. In this paper, such a problem is investigated for the task of automatic wine label recognition, and a novel data (3D viewpoint) augmentation technique is proposed to generate visually realistic training images, for essentially unlimited number of wine bottle poses, from a single wine label image captured in the real world. Experimental results show that the proposed augmentation technique can significantly improve the performance of the task of wine label recognition, by 13.8% over the traditional 2D image data augmentation, when the training data is extremely limited, e.g., having only one image for each wine class.

Acknowledgements. This work was supported by the Ministry of Science and Technology, Taiwan, under Contract NSTC 112-2634-F-A49 -007 and Contract NSTC 112-2221-E-A49 -060. This work was funded by ATEK Technology Corporation.

References

1. Vivino App: Scan any wine label to see which wines to buy and which to leave on the shelf (2023). https://www.vivino.com/app
2. Al-Qerem, A., Abu Salem, A., Jebreen, I., Nabot, A., Samhan, A.: Comparison between transfer learning and data augmentation on medical images classification. In: 22nd International Arab Conference on Information Technology (ACIT), pp. 1–7 (2021)
3. Mumuni, A., Mumuni, F.: Data augmentation: A comprehensive survey of modern approaches. Array **16**, 100258 (2022)
4. Alston, J.M., Gaeta, D.: Reflections on the political economy of European wine appellations. Italian Econ. J. **7**(2), 219–258 (2021)
5. Anaya-Isaza, A., Mera-Jiménez, L.: Data augmentation and transfer learning for brain tumor detection in magnetic resonance imaging. IEEE Access **10**, 23217–23233 (2022)
6. Angeli, A, et al.: Making paper labels smart for augmented wine recognition. In: Visual Computer (2023)
7. Baek, Y, et al: Character region awareness for text detection. In: CVPR, pp. 9365–9374 (2019)
8. Budach, L, et al.: The effects of data quality on machine learning performance. arXiv:2207.14529 (2022)
9. Caron, M., et al.: Emerging properties in self-supervised vision transformers. In: ICCV, pp. 9650–9660 (2021)
10. Chen, P.-Y., et al.: Mixed stage partial network and background data augmentation for surveillance object detection. IEEE Trans. Intell. Transpor. Syst. **23**(11), 23533–23547 (2022)
11. Cheng, C., Zhou, B., Ma, G., Wu, D., Yuan, Y.: Wasserstein distance based deep adversarial transfer learning for intelligent fault diagnosis with unlabeled or insufficient labeled data. Neurocomputing **409**(7), 35–45 (2020)
12. Cubuk, E.D., Zoph, B., Mane, D., Vasudevan, V., Le, Q.V.: Auto-augment: learning augmentation policies from data. In: CVPR, pp. 113–123 (2019)
13. Dosovitskiy, A., et al. N.: An image is worth 16 × 16 words: Transformers for image recognition at scale. In: ICLR, pp. 1–21 (2021)
14. Hermans, A, et al: In defense of the triplet loss for person re-identification. arXiv:1703.07737 (2017)

15. Jena, P.M., Tripathy, B.K., Parida, S.K.: Multi-font curved text scene for enhanced OCR readability. In: International Conference on Computational Intelligence and Data Engineering, pp. 223–235 (2022)
16. Kim, M., et al.: Deep learning in medical imaging. Neurospine **16**(4), 657–668 (2019)
17. Lamba, M., Madhusudhan, M.: Exploring OCR errors in full-text large documents: a study of LIS theses and dissertations. Lib. Philos. Pract. (e-journal), 7824 (2023)
18. Li, X., Ma, J.: Distributed search and fusion for wine label image retrieval. PeerJ Comput. Sci. **8**(e1116), 1–19 (2022)
19. Li, X., Yang, J., Ma, J.: CNN-SIFT consecutive searching and matching for wine label retrieval. In: ICIC 2019: Intelligent Computing Theories and Application, pp. 250–261 (2019)
20. Li, X., Yang, J., Ma, J.: Large scale category-structured image retrieval for object identification through supervised learning of CNN and SURF-based matching. IEEE Access **8**, 57796–57809 (2020)
21. Liao, M, et al.: Character region awareness for text detection. In: AAAI, p. 11474 (2020)
22. Liu, Y., Chen, H., Shen, C., He, T., Jin, L., Wang, L.: ABCNet: Real-time scene text spotting with adaptive Bezier-curve network. In: CVPR, pp. 9809–9818 (2020)
23. Long, S., Ruan, J., Zhang, W., He, X., Wu, W., Yao, C.: TextSnake: a flexible representation for detecting text of arbitrary shapes. In: ECCV, pp. 19–35 (2018)
24. Morger, A, et al: Studying and mitigating the effects of data drifts on ML model performance at the example of chemical toxicity data. In: Science Report (2022)
25. Rahman, A.B.M.A., et al.: Two decades of Bengali handwritten digit recognition: a survey. IEEE Access **10**, 92597–92632 (2022)
26. Schroff, F., Kalenichenko, D., Philbin, J.: FaceNet: a unified embedding for face recognition and clustering. In: CVPR, pp. 815–823 (2015)
27. Stacchio, L, et al: Rethinking augmented wine recognition. In: IEEE International Symposium on Mixed and Augmented Reality Adjunct (ISMAR-Adjunct), pp. 560–565 (2022)
28. Thakare, S., Kamble, A., Thengne, V., Kamble, U.R.: Document segmentation and language translation using tesseractocr. In: 13th International Conference on Industrial and Information Systems (ICIIS), pp. 148–151 (2018)
29. Vaswani, A, et al.: Attention is all you need. In: NeurIPS, pp. 5998–6008 (2017)
30. Wu, M.-Y.: Wine label image recognition using convolutional neural network with augmented data. In: International Conference on Computer Graphics, Visualization, Computer Vision and Image Processing (2019)
31. Huang, G., et al.: Learning to align from scratch. In: Neural Information Processing Systems, pp. 764–772 (2012)
32. Hassner, T, et al.: Effective face frontalization in unconstrained images (2015)
33. Taigman, Y, Yang, M., Ranzato, M.A., Wolf, L.: DeepFace: closing the gap to human-level performance in face verification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 1701–1708 (2014)
34. Bender, T., et al: Learning to taste: a multimodal wine dataset. In: Advances in Neural Information Processing Systems, vol. 36 (2024)

Lightweight Single Image Super-Resolution Network Integrating CNN and Transformer

Kai Zhu[1,2] and Li Chen[1,2](✉)

[1] School of Computer Science and Technology, Wuhan University of Science and Technology , Wuhan, China
chenli@wust.edu.cn

[2] Hubei Province Key Laboratory of Intelligent Information Processing and Real-time Industrial System, Wuhan University of Science and Technology, Wuhan, China

Abstract. Recently, remarkable progress has been achieved in single image super-resolution using methods based on CNN and Transformer architectures. However, existing approaches often construct a substantial number of network layers, leading to a significant increase in performance requirement and memory consumption, thereby limiting the practical deployment and usability of the models. To address this issue, we propose an Alternating CNN Transformer Block and an Integrative CNN Efficient Transformer for single image super-resolution. We enhance feature extraction efficiency by combining CNN within and between Transformer modules. In addition, we propose two novel structures: Multi-branch Gated CNN and Parallel Channel Attention, aiming to efficiently extract local spatial information and global channel information from images. Extensive experiments demonstrate that our model achieves high performance while maintaining low model complexity. The proposed model attains PSNR values of 32.32 and 30.78 for the Set5 and Manga109 benchmark datasets, respectively, at a scale factor of ×4. Compared to other lightweight super-resolution models, our proposed model outperforms them at lower computational costs. The source codes are available at https://github.com/kylechuuuuu/ICTSRN.

Keywords: SISR · CNN · Transformer · Attention · Deep Learning

1 Introduction

Single Image Super-Resolution (SISR) is an image processing technique employed to enhance the details and textures of images. Its objective is to enhance the clarity of low-resolution (LR) images by restoring blurred details, ultimately leading to a visually refined output that resembles high-resolution (HR) images. As a low-level visual task, SISR finds widespread applications in fields such as medical image enhancement [1,2], satellite image processing [3,4], video enhancement [5,6], and security surveillance [7,8]. Additionally, it

A. Antonacopoulos et al. (Eds.): ICPR 2024, LNCS 15332, pp. 212–225, 2025.
https://doi.org/10.1007/978-3-031-78125-4_15

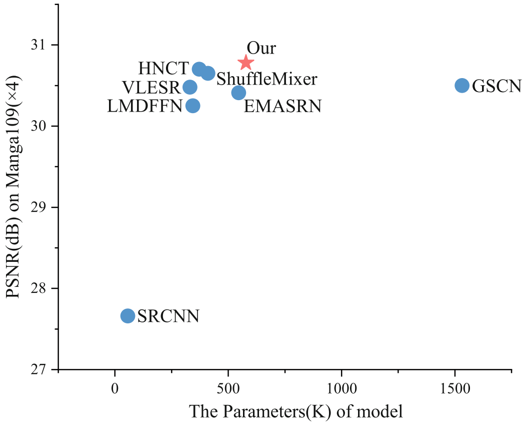


Fig. 1. Comparison of PSNR and Parameters on Manga109 (4×).

contributes to high-level vision tasks like object detection [9] and image segmentation [10].

In recent years, with the rapid development of hardware accelerators such as GPU and FPGA, deep learning-based SISR methods have experienced substantial progressions at a rapid pace. The inaugural efficacious endeavor employing CNN for SISR manifested in SRCNN [11]. Using a linear stack of convolution layers, SRCNN adeptly encapsulated the intricate interdependencies between LR and HR images, thereby achieving commendable results in the domain of high-quality image restoration. The work by Lim *et al.* [12] attained noteworthy performance enhancements by incorporating conventional residual modules into their methodology. Cheng *et al.* [13] introduced an encoder-decoder residual network specifically designed for efficient HR image restoration. Tian *et al.* [14] proposed a lightweight CNN-based SISR method by comprehensively integrating deep-channel and wide-channel features. However, to capture higher-level global feature information, CNN-based SISR models require incorporating deeper and larger network architectures. This gives rise to an undesirable escalation in computational complexity and hardware consumption, thereby presenting challenges in the deployment and utilization of the model. Additionally, purely CNN-based SISR methods lack competitiveness.

To enhance feature extraction efficiency and further reduce computational complexity, we propose an Alternating CNN-Transformer Block (ACTB), which alternately integrates CNN and Transformer modules to enhance feature extraction efficiency. Additionally, we propose a novel Integrative CNN Efficient Transformer (ICET). In ICET, we redesign Multi-branch Gated CNN (MGC) layer to obtain local spatial feature information and capture global channel feature information through Parallel Channel Attention (PCA). By combining local spatial information with global channel information, we achieve more efficient feature extraction, leading to excellent super-resolution performance with fewer computational resources. We compared ICTSRN with other recent lightweight SISR models in terms of PSNR and parameters, as shown in Fig. 1. From the figure, it can be observed that ICTSRN achieves high PSNR with low parameters.

The main contributions of our study can be summarized as follows:

- We propose a lightweight SISR network with lower computational cost, called ICTSRN. Compared with other existing lightweight models, our model has better performance at lower computational cost.
- To better integrate CNN and Transformer, we propose ICET, in which we redesign and introduce two novel components: MGC and PCA, to extract local spatial information and global channel information.
- We propose the ACTB module, which achieves a simple and efficient integration of CNN and Transformer modules by alternately stacking them and incorporating residual connections.

2 Related Work

2.1 CNN-Based SISR Method

Early CNN-based SISR methods, such as SRCNN [11] and VDSR [15], demonstrated the huge potential of deep learning in SISR. These pioneering studies established a foundational framework for subsequent research, demonstrating significant advancements over traditional non-deep learning approaches. Recent years have seen the development of more sophisticated CNN architectures. Notable examples include EDSR [12], which leveraged residual learning, and RCAN [16], which introduced channel attention mechanisms. Mei *et al.* [17] proposed a novel non-local sparse attention mechanism for SISR by combining non-local operations with sparse representation. Addressing the need for efficient SISR methods, especially for resource constrained devices, several lightweight architectures have been proposed. FALSR [18] employed neural architecture search to find efficient network designs, and IMDN [19] focused on information multi-distillation to reduce model size while maintaining performance.

2.2 Transformer-Based SISR Method

Transformer architectures have recently made significant advancements into computer vision tasks, including SISR. These models have demonstrated remarkable performance, often outperforming conventional CNN-based approaches in quality and efficiency. IPT [20] pioneered this approach, utilizing a pre-trained model and fine-tuning strategy to achieve remarkable performance in SISR tasks. SwinIR [21] adapted the Swin Transformer architecture for image restoration, introducing shifted windows to efficiently capture both local and global dependencies, thus striking a balance between computational cost and restoration quality. HAT [22] proposed to enhance image reconstruction by integrating channel attention and window-based self-attention mechanisms, leveraging their complementary strengths: the utilization of global statistical information and powerful local fitting capabilities. Another notable contribution came from the ESRT [23], which focused on efficiency while maintaining high performance by employing a progressive learning strategy and introducing an efficient multi-scale self-attention mechanism, significantly reducing computational complexity compared to previous Transformer-based models.

2.3 Lightweight CNN-Transformer Fusion for SISR

SISR methods leveraging Transformer have exhibited noteworthy performance. Nevertheless, the computational complexity of the Vanilla Transformer is excessively high, rendering it unsuitable for direct integration into the SISR domain. Furthermore, the Transformer model encounters challenges related to its limited capacity for extracting local detailed features. Therefore, a viable solution involves combining it with CNN. HNCT [24] encapsulated the CNN and Transformer into a block to achieve both local and global feature extraction. LBNet [25] innovatively integrates a Symmetric CNN for local feature extraction with a Recursive Transformer to capture long-term dependencies, culminating in a lightweight and efficient approach for SISR. Yoo *et al.* [26] proposed a super-resolution network composed of a parallel fusion of CNN and Transformer backbone. Liu *et al.* [27] proposed a lightweight super-resolution network composed of a Transformer cluster and a CNN module cluster. However, existing methods often fail to achieve an optimal integration of the local feature extraction capabilities of CNN and the intrinsic attention mechanisms of Transformer. Furthermore, the substantial complexity of these models, in terms of both parameter count and FLOPs, remains a significant consideration.

3 The Proposed Method

3.1 Overall Structure

The main structure of the proposed ICTSRN is illustrated in Fig. 2, generally divided into three parts: a basic feature extraction module consisting of a 3×3 convolution layer, a deep feature extraction module composed of ACTB concatenation and an image reconstruction module consisting of a 3×3 convolution and a pixel-shuffle layer. In the basic feature extraction stage, we employ a 3×3 convolution to capture fundamental features from the LR image $X_{LR} \in \mathbb{R}^{3\times H\times W}$ and expand the channels

$$X_B = H_{SF}(X_{LR}), \tag{1}$$

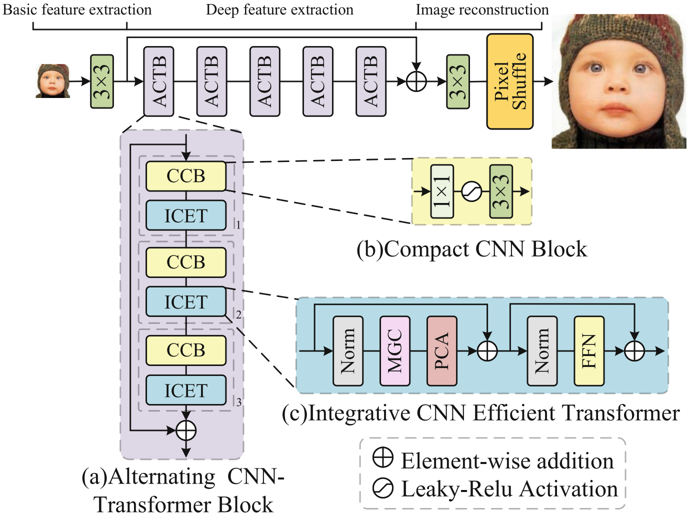


Fig. 2. The main structure of the ICTSRN, ACTB, CCB and ICET.

where H_{SF} denotes the basic feature extraction module and X_B denotes the extracted basic feature. Subsequently, we utilize a deep feature extraction module, which is constructed by concatenating the ACTB blocks, to extract high-level feature information

$$H_D = H^n_{ACTB}(H^{n-1}_{ACTB} \cdots (H^1_{ACTB}(X_B))), \tag{2}$$

where H^n_{ACTB} represents the n-th ACTB module and H_D represents the high-level features output from the n-th ACTB module. Finally, the integration of H_B and H_D is fed into the reconstruction module to generate a HR image I_{SR}

$$I_{HR} = H_{RC}(H_B + H_D), \tag{3}$$

where H_{RC} represents the reconstruction module, and $I_{HR} \in \mathbb{R}^{3\times H\times W}$ represents the reconstructed HR image.

We utilize the L1 loss function for training our model due to its efficient computation and robustness against outliers, which collectively enhance the model's generalization capabilities. The loss function of ICTSRN can be expressed as follows:

$$LOSS = \frac{1}{N}\sum_{i=1}^{N} ||H_{ICTSRN}\left(L^i_{LR}\right) - I^i_{HR}||_1, \tag{4}$$

where H_{ICTSRN} represents the ICTSRN, $\|\cdot\|_1$ represents the L_1 norm, N is the number of training samples, I^i_{LR} and I^i_{HR} represent the i-th input LR image and its corresponding HR image respectively.

3.2 Alternative CNN Transformer Block

The objective of the ACTB design is to enhance the integration between CNN and Transformer to improve the efficiency of feature extraction. ACTB consists of three ICET and three Compact CNN Block (CCB) blocks, which are alternately arranged as shown in Fig. 2(a). The primary advantage of the alternating architectural design resides in its capacity to synergistically harness the unique strengths of convolutional layers and Transformers. This strategy facilitates an enhanced extraction and comprehensive analysis of both local and global feature information from the input imagery, thereby augmenting the model's feature extraction capability. The CCB module consists of a 1×1 convolution, Leaky-ReLU activation function and a 3×3 convolution. The structure of ICET will be explained in detail in Sect. 3.3. For the n-th input F_{in}^n of ACTB, the function of ACTB can be described as follows:

$$\begin{aligned} F_{out}^n &= H_{ACTB}^n(F_{in}^n) + F_{in}^n \\ &= H_{CCB}^{n,3}(H_{ICET}^{n,3}(\cdots(H_{CCB}^{n,1}(H_{ICET}^{n,1}(F_{in}^n))))) + F_{in}^n \end{aligned} \tag{5}$$

where H_{ACTB}^n and F_{out}^n represents the n-th ACTB and the output; $H_{CCB}^{n,i}$ and $H_{ICET}^{n,i}$ represents the i-th CCB and ICET in the n-th ACTB module.

3.3 Integrative CNN Efficient Transformer

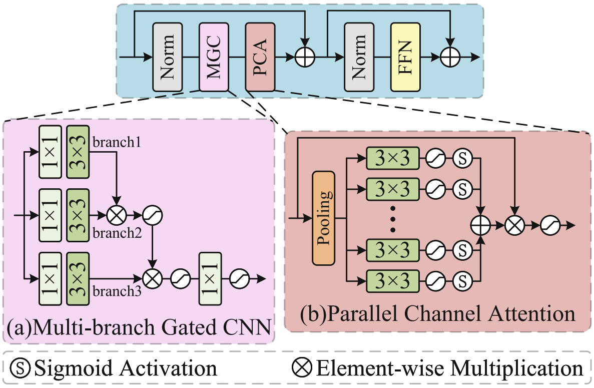

Norm
MGC
PCA
Norm
FFN
1×1
3×3
branch1
branch2
branch3
(a)Multi-branch Gated CNN
Pooling
3×3
(b)Parallel Channel Attention
Ⓢ Sigmoid Activation
⊗ Element-wise Multiplication

Fig. 3. The structure of the MGC and PCA in ICET.

The Transformer exhibits a strong capability for capturing global information and global receptive field. However, it lacks local feature extraction ability in comparison to CNN. To address this issue, we propose ICET. As illustrated in Fig. 3, the self-attention in the Vanilla Transformer is replaced with the proposed MGC and PCA. Starting with the MGC, the input features' channels are initially expanded to three times their original size using a 1×1 convolution, as shown in Fig. 3(a). Following the expansion, the augmented channels are partitioned into three equal segments along the channel axis. And local spatial information is obtained through a 3×3 convolution. Subsequently, branch 1 is multiplied by branch 2 to form a gating mechanism; the product undergoes an activation function and is multiplied by branch 3, forming another gating mechanism. Finally, feature aggregation is accomplished through a 1×1 convolution, and the result is output after passing through a Leaky-ReLU activation function. For the module input F_{in}^{MGC}, MGC is formulated as:

$$H_{mid}^{MGC} = \phi(C_1^1(C_3^1(F_{in}^{MGC})) \otimes C_1^2(C_3^2(F_{in}^{MGC}))), \tag{6}$$

$$H_{MGC} = \phi(C_1^4(H_{mid}^{MGC} \otimes C_1^3(C_3^3(F_{in}^{MGC})))), \tag{7}$$

where H_{MGC} represents MGC; C_1^i, C_3^i, ϕ and $\otimes$ represent the i-th 1×1 convolution, i-th 3×3 convolution, Leaky-ReLU activation and element-wise multiplication respectively.

The attention has the ability to extract global features, which can complement CNN with smaller receptive fields and achieve more detailed feature extraction. However, self-attention is computationally expensive and not suitable for lightweight models. Therefore, we use channel attention for feature extraction. Unlike traditional channel attention mechanisms, we were inspired by parallel mechanism to propose PCA. The object of PCA is to enhance valuable channel informations by summing up the multiple parallel branches. As illustrated in Fig. 3(b), given an input F_{in}^{PCA}, PCA first reshapes the input through global pooling, compressing global spatial information. To reduce parameter overhead, we use group convolution instead of setting a reduction ratio. Additionally, to minimize the loss of information during dimensionality reduction, each branch is equipped with a single layer of 3×3 convolution using group convolution. Subsequently, the weights from different parallel branches are aggregated by summation and then multiplied with the input F_{in}^{PCA}. Ultimately, the output is obtained after passing it through Leaky-ReLU function. To summarize, the function of PCA can be outlined as follows:

$$H_{PCA} = \phi(F_{in}^{PCA} \otimes \sum_{i=1}^{j} \sigma(\phi(C_3^j(P(F_{in}^{PCA}))))), \tag{8}$$

where H_{PCA} represents the function of PCA; C_3^j represents the j-th 3×3 convolution; σ and P represent the Sigmoid activation and Pooling respectively. Finally, the Feed-Forward Network (FFN) in Transformer, which has the same structure as CCB, is used to perform feature transformation and generate output. Given the input F_{in}^{ICET}, the ICET can be formulated as follow:

$$H_{mid}^{ICET} = H_{PCA}(H_{MGC}(Norm(F_{in}^{ICET}))) + F_{in}^{ICET}, \tag{9}$$

$$H_{ICET} = FFN(Norm(H_{mid}^{ICET})) + H_{mid}^{ICET}, \tag{10}$$

where H_{ICET} represents the function of ICET; $Norm$ represents Layer Normalization and FFN represents a Feed-Forward Network.

4 Experiments

4.1 Experimental Setup

We utilized the DF2K [28] for both training and validation and employed five benchmark datasets: Set5 [29], Set14 [30], BSD100 [31], Urban100 [32] and Manga109 [33] as our test sets. Our ICTSRN architecture incorporated 5 ACTB modules, while the PCA component was designed with 12 branches. Additionally, we configured the channel count to be expanded to 64. All experiments were executed on a computer running the Ubuntu 20.04 operating system with a NVIDIA RTX A5000 24G GPU. The model architecture was designed and implemented leveraging the PyTorch framework. Model training involved minimizing the loss using the Adam optimizer, with $\beta_1 = 0.9$ and $\beta_2 = 0.999$, over a total of 900,000 iterations. The initial learning rate was set to 5×10^{-4}, and a cosine annealing scheduler was employed to reduce it to 5×10^{-6}. The training batch size was set to 48, with a patch size of 256×256. Data augmentation techniques, including random rotation and random horizontal flipping, were applied during training.

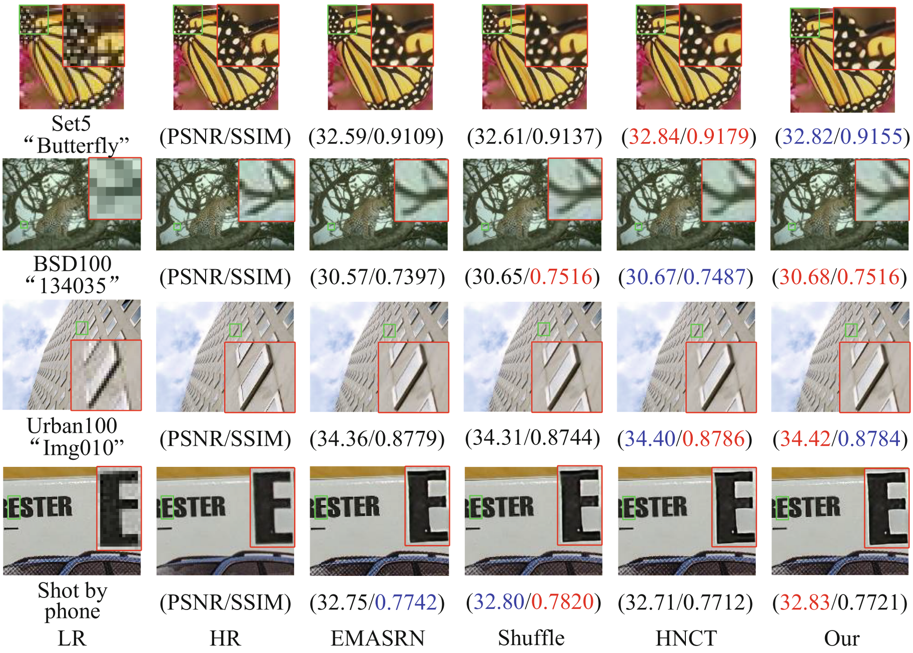


Fig. 4. Qualitative comparison for ×4 upscaling in four pictures.

Table 1. Model comparison results with different lightweight sisr methods, red represents the best performance, while blue indicates the second best.

Method	Scale	Set5 [29]	Set14 [30]	BSD100 [31]	Urban100 [32]	Manga109 [33]
		PSNR/SSIM	PSNR/SSIM	PSNR/SSIM	PSNR/SSIM	PSNR/SSIM
Bicubic	×3	30.39/0.8682	27.55/0.7742	27.21/0.7385	24.46/0.7349	26.95/0.8556
SRCNN [11]		32.75/0.9090	29.28/0.8209	28.41/0.7863	26.24/0.7989	30.59/0.9107
GSCN [34]		34.40/0.9271	30.35/0.8425	29.11/0.8053	28.20/0.8535	33.54/0.9445
EMASRN [35]		34.36/0.9264	30.30/0.8411	29.05/0.8035	28.04/0.8493	33.43/0.9433
ShuffleMixer [36]		34.40/0.9272	30.37/0.8423	29.12/0.8051	28.08/0.8498	33.69/0.9448
HNCT [24]		34.47/0.9275	30.44/0.8439	29.15/0.8067	28.28/0.8557	33.81/0.9459
ACDN [37]		34.39/0.9262	30.32/0.8419	29.12/0.8053	28.26/0.8542	-
VLESR [38]		34.40/0.9272	30.34/0.8415	29.08/0.8043	28.16/0.8519	33.61/0.9445
LMDFFN [39]		34.32/0.9264	30.20/0.8392	29.03/0.8034	28.01/0.8483	33.36/0.9430
Our		34.46/0.9278	30.40/0.8430	29.13/0.8058	28.30/0.8553	33.76/0.9457
Bicubic	×4	28.42/0.8104	26.00/0.7027	25.96/0.6675	23.14/0.6577	24.89/0.7866
SRCNN [11]		30.48/0.8628	27.49/0.7503	26.90/0.7101	24.52/0.7221	27.66/0.8505
GSCN [34]		32.18/0.8950	28.60/0.7821	27.59/0.7364	26.12/0.7872	30.50/0.9080
EMASRN [35]		32.17/0.8948	28.57/0.7809	27.55/0.7351	26.01/0.7838	30.41/0.9076
ShuffleMixer [36]		32.21/0.8953	28.66/0.7827	27.61/0.7366	26.08/0.7835	30.65/0.9093
HNCT [24]		32.31/0.8957	28.71/0.7834	27.63/0.7381	26.20/0.7896	30.70/0.9112
ACDN [37]		32.30/0.8950	28.64/0.7819	27.59/0.7361	26.22/0.7891	-
VLESR [38]		32.17/0.8945	28.55/0.7802	27.55/0.7345	26.03/0.7830	30.48/0.9073
LMDFFN [39]		32.08/0.8930	28.46/0.7792	27.51/0.7341	25.93/0.7804	30.25/0.9053
Our		32.32/0.8967	28.70/0.7838	27.64/0.7384	26.25/0.7909	30.78/0.9118

Table 2. Model complexity comparisons for ×4 scale factor. The assessment of model Params, FLOPs, and Runtime is conducted using input images sized at 256×256.

Method	EMASRN [35]	ShuffleMixer [36]	HNCT [24]	VLESR [38]	Our
Params	546K	411K	372K	331K	578K
FLOPs	480.3G	17.9G	39.4G	19.5G	26.9G
Runtime	70.1 ms	24.7 ms	325 ms	47 ms	41.1 ms

4.2 Experimental Results

Quantitative Evaluation. Based on the experimental results presented in Table 1, we conducted a comprehensive evaluation of our proposed lightweight SISR model. Our model was compared with several state-of-the-art methods across five widely used benchmark datasets (Set5, Set14, BSD100, Urban100, and Manga109). These competing methods include CNN-based structures such as SRCNN [11], ShuffleMixer [36], ACDN [37], and VLESR [38], attention-based mechanisms like EMASRN [35] and GSCN [34], and the hybrid CNN-Transformer structure HNCT [24]. The results demonstrate that at ×3 scale, our model exhibits excellent performance on most datasets, typically ranking second or third. At ×4 scale, our model's performance is even more impressive, achieving the best or second-best results across all datasets. Specifically, for the ×4 scale, our model attains the highest PSNR and SSIM scores on Set5, BSD100, Urban100, and Manga109 datasets. These findings strongly validate the effectiveness and advanced nature of our proposed method, particularly for the more challenging ×4 scale SR task. Notably, as a lightweight approach, our model achieves such outstanding performance while maintaining low computational complexity, indicating a favorable balance between model efficiency and SR quality. As shown in Table 2, our model outperforms several existing methods in terms of both accuracy and computational efficiency. Specifically, our model has 578K parameters, 26.9G FLOPs, and a runtime of 41.1 ms, which is a notable improvement over the compared methods. This result has significant implications for resource-constrained scenarios in practical applications.

Qualitative Analysis. We selected images from Set5, BSD100, Urban100, and a smartphone-captured photo for qualitative experimentation. As shown in Fig. 4, our model excels in reconstructing texture details compared to other models. In the "Butterfly" image from Set5, our model achieves a PSNR of 32.82 and an SSIM of 0.9155, showing superior detail restoration. For the BSD100 "134035" image, our model attains a PSNR of 30.68 and an SSIM of 0.7516, maintaining clarity in complex textures. The Urban100 "Img010" image demonstrates our model's capability with a PSNR of 34.42 and an SSIM of 0.8784, effectively handling urban textures. In the smartphone-captured image, our method achieves a PSNR of 32.83 and an SSIM of 0.7721, outperforming other models in retaining fine textural details. These results confirm our model's superior performance in practical applications.

4.3 Ablation Study

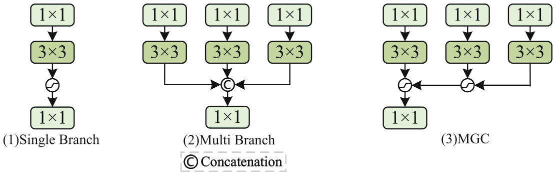
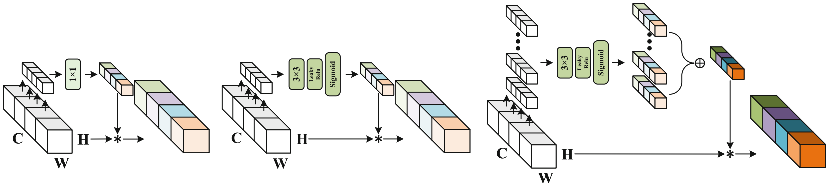


Fig. 5. Single Branch, Multi Branch and Multi-Branch Gate in ablation study.

Ablation Study of the MGC. To validate the effectiveness of the MGC, we designed two additional modules as replacements. Their structures are illustrated in Fig. 5, featuring a single-branch structure, a multi-branch structure, and our proposed MGC. As shown in Table 3, experimental results on the Set5 and BSD100 datasets demonstrate that our proposed MGC enhances performance without significantly increasing computational costs.



(1)Simple Channel Attention (2)Channel Attention (3)PCA

Fig. 6. SCA, CA and PCA in ablation study.

Ablation Study of the PCA. As depicted in Fig. 6, we also designed two additional modules, Simple Channel Attention and Channel Attention, to replace PCA in the ICET and validate its effectiveness. The results, shown in Table 3, demonstrate improved PSNR and SSIM on the Set5 and BSD100 datasets without increasing computational costs.

Table 3. Ablation study of the MGC and PCA.

Model	Params	FLOPs	Runtime	Set5	BSD100
Single	484K	20.5G	43 ms	32.17/0.8946	27.39/0.7347
Multi	669K	33.9G	46 ms	32.26/0.8963	27.58/0.7373
MGC	578K	26.9G	41.1 ms	32.32/0.8967	27.64/0.7384
SCA	540K	26.9G	28.6 ms	32.19/0.8949	27.43/0.7352
CA	484K	26.9G	32.6 ms	32.25/0.8957	27.56/0.7364
PCA	578K	26.9G	41.1 ms	32.32/0.8967	27.64/0.7384

5 Conclusion

In this paper, we propose ACTB, MGC and PCA to construct a lightweight super-resolution network ICTSRN. To achieve better results, we combine the advantages of both CNN and Transformer. Extensive experiments have demonstrated the effectiveness of our approach. However, the performance of our model in ×3 super-resolution is not outstanding, and there is overfitting in the model. In the future, we will further improve the efficiency of feature extraction in PCA to enhance the performance of the model under ×3 scale and reduce computational costs.

Acknowledgements. This work was supported by the National Natural Science Foundation of China (62271359).

References

1. Ren, S., Guo, K., Ma, J., Zhu, F., Hu, B., Zhou, H.: Realistic medical image super-resolution with pyramidal feature multi-distillation networks for intelligent healthcare systems. Neural Comput. Appl. **35**, 22781–22796 (2023)
2. Dharejo, F., et al.: Multimodal-boost: multimodal medical image super-resolution using multi-attention network with wavelet transform. IEEE/ACM Trans. Comput. Biol. Bioinform. (2022)
3. Xiao, Y., Yuan, Q., Jiang, K., He, J., Wang, Y., Zhang, L.: From degrade to upgrade: learning a self-supervised degradation guided adaptive network for blind remote sensing image super-resolution. Inf. Fusion. **96**, 297–311 (2023)
4. Xiao, Y., et al.: Local-global temporal difference learning for satellite video super-resolution. ArXiv Preprint ArXiv:2304.04421. (2023)
5. Yi, P., Wang, Z., Jiang, K., Jiang, J., Lu, T., Ma, J.: A progressive fusion generative adversarial network for realistic and consistent video super-resolution. IEEE Trans. Pattern Anal. Mach. Intell. **44**, 2264–2280 (2020)
6. Lu, Y., Wang, Z., Liu, M., Wang, H., Wang, L.: Learning spatial-temporal implicit neural representations for event-guided video super-resolution. In: Proceedings of the IEEE/CVF Conference on Computer Vision And Pattern Recognition, pp. 1557–1567 (2023)
7. Jiang, K., Wang, Z., Yi, P., Wang, G., Gu, K., Jiang, J.: ATMFN: adaptive-threshold-based multi-model fusion network for compressed face hallucination. IEEE Trans. Multimedia **22**, 2734–2747 (2019)
8. Jiang, K., Wang, Z., Yi, P., Lu, T., Jiang, J., Xiong, Z.: Dual-path deep fusion network for face image hallucination. IEEE Trans. Neural Netw. Learn. Syst. **33**, 378–391 (2020)
9. Zhang, J., Lei, J., Xie, W., Fang, Z., Li, Y., Du, Q.: SuperYOLO: super resolution assisted object detection in multimodal remote sensing imagery. IEEE Trans. Geosci. Remote Sens. **61**, 1–15 (2023)
10. Chen, S., Ogawa, Y., Zhao, C., Sekimoto, Y.: Large-scale individual building extraction from open-source satellite imagery via super-resolution-based instance segmentation approach. ISPRS J. Photogrammetry Remote Sens. **195**, 129–152 (2023)
11. Dong, C., Loy, C., He, K., Tang, X.: Learning a deep convolutional network for image super-resolution. In: Computer Vision-ECCV 2014: 13th European Conference, Zurich, Switzerland, 6-12 September 2014, Proceedings, Part IV 13, pp. 184–199 (2014)
12. 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)
13. Cheng, G., Matsune, A., Li, Q., Zhu, L., Zang, H., Zhan, S.: Encoder-decoder residual network for real super-resolution. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (2019)
14. Tian, C., Yuan, Y., Zhang, S., Lin, C., Zuo, W., Zhang, D.: Image super-resolution with an enhanced group convolutional neural network. Neural Netw. **153**, 373–385 (2022)

15. Kim, J., Lee, J., Lee, K.: 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. 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)
17. 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)
18. Ahn, N., Kang, B., Sohn, K.: Fast, accurate, and lightweight super-resolution with cascading residual network. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 252–268 (2018)
19. 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)
20. Chen, H., et al.: Pre-trained image processing transformer. In: Proceedings of the IEEE/CVF Conference on Computer Vision And Pattern Recognition, pp. 12299–12310 (2021)
21. 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)
22. 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)
23. 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)
24. Fang, J., Lin, H., Chen, X., Zeng, K.: A hybrid network of CNN and transformer for lightweight image super-resolution. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 1103–1112 (2022)
25. Gao, G., Wang, Z., Li, J., Li, W., Yu, Y., Zeng, T.: Lightweight bimodal network for single-image super-resolution via symmetric CNN and recursive transformer. ArXiv Preprint ArXiv:2204.13286. (2022)
26. Yoo, J., Kim, T., Lee, S., Kim, S., Lee, H., Kim, T.: Enriched CNN-transformer feature aggregation networks for super-resolution. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pp. 4956–4965 (2023)
27. Liu, Y., Yue, M., Yan, H., Zhu, L.: Single-image super-resolution using lightweight transformer-convolutional neural network hybrid model. IET Image Process. (2023)
28. 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)
29. Bevilacqua, M., Roumy, A., Guillemot, C., Alberi-Morel, M.: Low-complexity single-image super-resolution based on nonnegative neighbor embedding. In: Proceedings of the British Machine Vision Conference, pp. 135.1–135.10 (2012)
30. Yang, J., Wright, J., Huang, T., Ma, Y.: Image super-resolution via sparse representation. IEEE Trans. Image Process. **19**, 2861–2873 (2010)
31. 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 (2001)

32. Huang, J., 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)
33. Matsui, Y., et al.: Sketch-based manga retrieval using manga109 dataset. Multimedia Tools Appl. **76**, 21811–21838 (2017)
34. Liu, C., Lei, P.: An efficient group skip-connecting network for image super-resolution. Knowl. Based Syst. **222**, 107017 (2021)
35. Zhu, X., Guo, K., Ren, S., Hu, B., Hu, M., Fang, H.: Lightweight image super-resolution with expectation-maximization attention mechanism. IEEE Trans. Circ. Syst. Video Technol. **32**, 1273–1284 (2021)
36. Sun, L., Pan, J., Tang, J.: ShuffleMixer: an efficient ConvNet for image super-resolution. Adv. Neural. Inf. Process. Syst. **35**, 17314–17326 (2022)
37. Wu, J., Wang, Y., Zhang, X.: Lightweight asymmetric convolutional distillation network for single image super-resolution. IEEE Sig. Process. Lett. (2023)
38. Gao, D., Zhou, D.: A very lightweight and efficient image super-resolution network. Expert Syst. Appl. **213**, 118898 (2023)
39. Guo, X., Tu, Z., Li, G., Shen, Z., Wu, W.: A novel lightweight multi-dimension feature fusion network for single-image super-resolution reconstruction. Vis. Comput. **40**, 1685–1696 (2024)

A Synthetic Benchmarking Pipeline to Compare Camera Calibration Algorithms

Lala Shakti Swarup Ray[1](✉), Bo Zhou[1,2], Lars Krupp[1,2], Sungho Suh[1,2], and Paul Lukowicz[1,2]

[1] German Research Center for Artificial Intelligence,Kaiserslautern, Germany
lala_shakti_swarup.ray@dfki.de
[2] RPTU Kaiserslautern-Landau,Kaiserslautern, Germany

Abstract. Accurate camera calibration is crucial for various computer vision applications. However, measuring calibration accuracy in the real world is challenging due to the lack of datasets with ground truth to evaluate them. In this paper, we present SynthCal, a synthetic camera calibration benchmarking pipeline that generates images of calibration patterns to measure and enable accurate quantification of calibration algorithm performance in camera parameter estimation. We present a SynthCal-generated calibration dataset with four common patterns, two camera types, and two environments with varying view, distortion, lighting, and noise levels for both monocular and multi-camera systems. The dataset evaluates both single and multi-view calibration algorithms by measuring re-projection and root-mean-square errors for identical patterns and camera settings. Additionally, we analyze the significance of different patterns using different calibration configurations. The experimental results demonstrate the effectiveness of SynthCal in evaluating various calibration algorithms and patterns.

Keywords: camera calibration · benchmarking · synthetic dataset · pattern recognition

1 Introduction

When we capture an image using a camera, the captured digital image can differ from the real-world scene in terms of perspective, distortion, color, resolution, and other visual properties. This is because real-world scenes are three-dimensional and continuous, while digital images captured by a camera are two-dimensional and discrete, and contain distortion and other imperfections. To minimize these differences and improve the accuracy of image-based computer vision tasks, camera calibration is essential.

Camera calibration involves calculating camera parameters that refer to its intrinsic and extrinsic characteristics for accurately mapping points in the 3D

A. Antonacopoulos et al. (Eds.): ICPR 2024, LNCS 15332, pp. 226–238, 2025.
https://doi.org/10.1007/978-3-031-78125-4_16

world to their corresponding 2D image coordinates. Once the camera is calibrated, it can accurately measure distances, angles, and sizes of objects in the 3D world and perform other image-based computer vision tasks such as object tracking [21], 3D reconstruction [12,20], medical imaging [2], and autonomous driving [8].

Geometric camera calibration [11,13] is one of the most widely used calibration methods. It involves using a calibration target with known geometric features, such as a calibration grid, to estimate the camera parameters.

However, creating real camera calibration data with ground truth for calibration algorithms can be challenging because it is difficult to measure camera position and rotation accurately, and the camera's intrinsic parameters can change with changes in the zoom level, focus distance, or temperature. Moreover, cameras can have different intrinsic parameters, even if they are of the same make and model, because of manufacturing tolerances, assembly errors, and differences in lens quality. Observing the calibration pattern in the image along with the previous knowledge of the pattern, we can determine the intrinsic and extrinsic parameters using various calibration algorithms, such as Zhang [22], Tsai [17], or Bouguet calibration method [3]. Previous works have tried to compare different camera calibration algorithms [19]. However, there is a need for a benchmarking procedure that can provide a quantitative comparison of calibration algorithms due to the unknown ground truth of the calibration dataset.

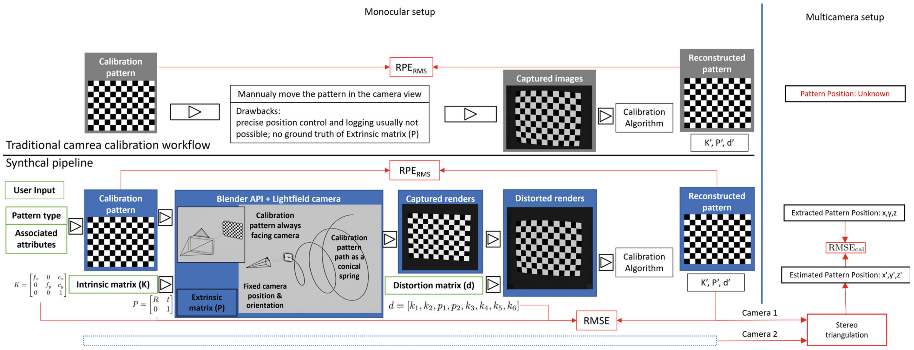

Monocular setup
Multicamera setup
Calibration pattern
RPE$_{RMS}$
Mannualy move the pattern in the camera view
Drawbacks:
precise position control and logging usually not possible; no ground truth of Extrinsic matrix (P)
Captured images
Calibration Algorithm
Reconstructed pattern
K', P', d'
Pattern Position: Unknown
Traditional camrea calibration workflow
Synthcal pipeline
RPE$_{RMS}$
User Input
Pattern type
Associated attributes
Calibration pattern
Blender API + Lightfield camera
Calibration pattern always facing camera
Calibration pattern path as a conical spring
Fixed camera position & orientation
Extrinsic matrix (P)
$K = \begin{bmatrix} f_x & 0 & c_x \\ 0 & f_y & c_y \\ 0 & 0 & 1 \end{bmatrix}$
Intrinsic matrix (K)
$P = \begin{bmatrix} R & t \\ 0 & 1 \end{bmatrix}$
Captured renders
Distorted renders
Distortion matrix (d)
$d = [k_1, k_2, p_1, p_2, k_3, k_4, k_5, k_6]$
Calibration Algorithm
Reconstructed pattern
K', P', d'
RMSE
Extracted Pattern Position: x,y,z
RMSE$_{cal}$
Estimated Pattern Position: x',y',z'
Camera 1
Camera 2
Stereo triangulation

Fig. 1. SynthCal pipeline to generate calibration dataset from a set of input attributes: Pattern attributes, camera intrinsic, distortion, extrinsic matrix. The accuracy is then evaluated using $RMSE$ and RPE_{RMS} for monocular cameras.

To overcome these problems, we introduce an overall pipeline, named SynthCal, which generates a synthetic camera calibration dataset with user-defined intrinsic camera parameters while precisely measuring the extrinsic camera parameters. It enables the selection of the optimal camera calibration algorithm for specific configurations by considering all intrinsic, extrinsic, and distortion parameters. Additionally, it ensures that lighting conditions and noise are identical for the different captured datasets for accurate comparison of different calibration patterns which is not possible in the real world. The idea of generating synthetic calibration data has been previously applied in other works, such as sports-based synthetic calibration [5] and evaluating closed-form solutions of principal line calibration [6], but not necessarily for comparing calibration algorithms.

Our main contributions can be summarized as follows:

- We present a pipeline to generate a camera calibration dataset with ground truth parameters and select the optimal camera calibration algorithm for the specific configurations, as depicted in Fig. 1.
- We validate the proposed pipeline on three different camera calibration algorithms which is consistent with previous works and then use SynthCal-generated dataset to compare four different calibration patterns given in Fig. 2 for monocular and multi-camera systems with two distinct camera configurations, and two different lighting and noise conditions.

2 Proposed Method

We created a modular web-based interface with OpenCV and Blender API in the back-end to generate a synthetic camera calibration dataset with ground truth which has functionalities to create different camera calibration patterns, simulate a camera inside Blender using the light-field analysis add-on [10], render the camera calibration pattern from various positions and orientations, add radial distortions while establishing the camera's intrinsic, extrinsic and distortion parameters to formulate the ground truth. We used an OpenCV to generate geometric patterns that take input pattern type, and pattern attributes to generate a PNG image. Our script allows us to create checkerboard patterns (Ch), symmetric circular patterns (Sc), asymmetric circular patterns (Ac), and Charuco [1] patterns (Cu) of different configurations as shown in Fig. 2. Let K be the intrinsic matrix of the camera, which includes the parameters that describe the internal configuration of the camera, such as the focal length (f_x, f_y) and principal point (c_x, c_y):

$$K = \begin{bmatrix} f_x & 0 & c_x \\ 0 & f_y & c_y \\ 0 & 0 & 1 \end{bmatrix} \quad (1)$$

We used a Blender python API and a light-field add-on to create synthetic cameras that take camera attributes (f_x, f_y) and (c_x, c_y) to create a camera configuration file for simulating the camera inside Blender. To capture the calibration pattern for dataset creation, we moved the pattern in a path resembling the shape of a conical spring, as depicted in Fig. 1. The center of the calibration pattern is always in the camera's direction, so the planar pattern can be captured in different angles, sizes, and orientations and have consistency without going out of the camera frame. Let R be the rotation matrix that describes the orientation of the camera in the global coordinate system, and let t be the translation vector that describes the position of the camera in the world coordinate system:

$$P = \begin{bmatrix} R & t \\ 0 & 1 \end{bmatrix} \tag{2}$$

The extrinsic matrix P combines the rotation matrix and the translation vector. R is, and t are evaluated by extracting the global position and orientation of the camera and calibration pattern at each frame. The camera parameters can also be described using the distortion parameters, which describe the deviations from the ideal imaging system. The distortion parameters can be represented as a vector

$$d = [k_1, k_2, p_1, p_2, k_3, k_4, k_5, k_6] \tag{3}$$

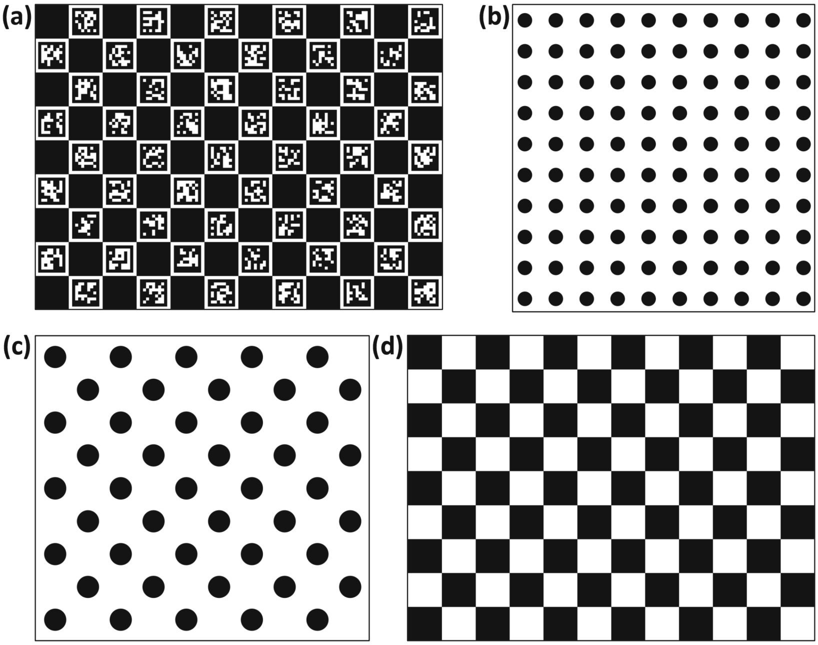


Fig. 2. (a) 9 × 12 Charuco pattern, (b) 10 × 10 Symmetric circle grid, (c) 9 × 10 Asymmetric circle grid, (d) 9 × 12 Checkerboard pattern.

where $k_1, k_2, k_3, k_4, k_5, k_6$ are radial distortion coefficients and p_1, p_2 are tangential distortion coefficients. The distortions are added later using Blender undistorted node by setting up a tracking scene in Blender and defining K and d. The final equation for mapping X a 3D point in the global coordinate system to x a 2D point in the image plane and s scale factor, including the distortion parameters, can be written as:

$$sx = K[R \mid t]X + d(\frac{x_d}{f_x}, \frac{y_d}{f_y}) \tag{4}$$

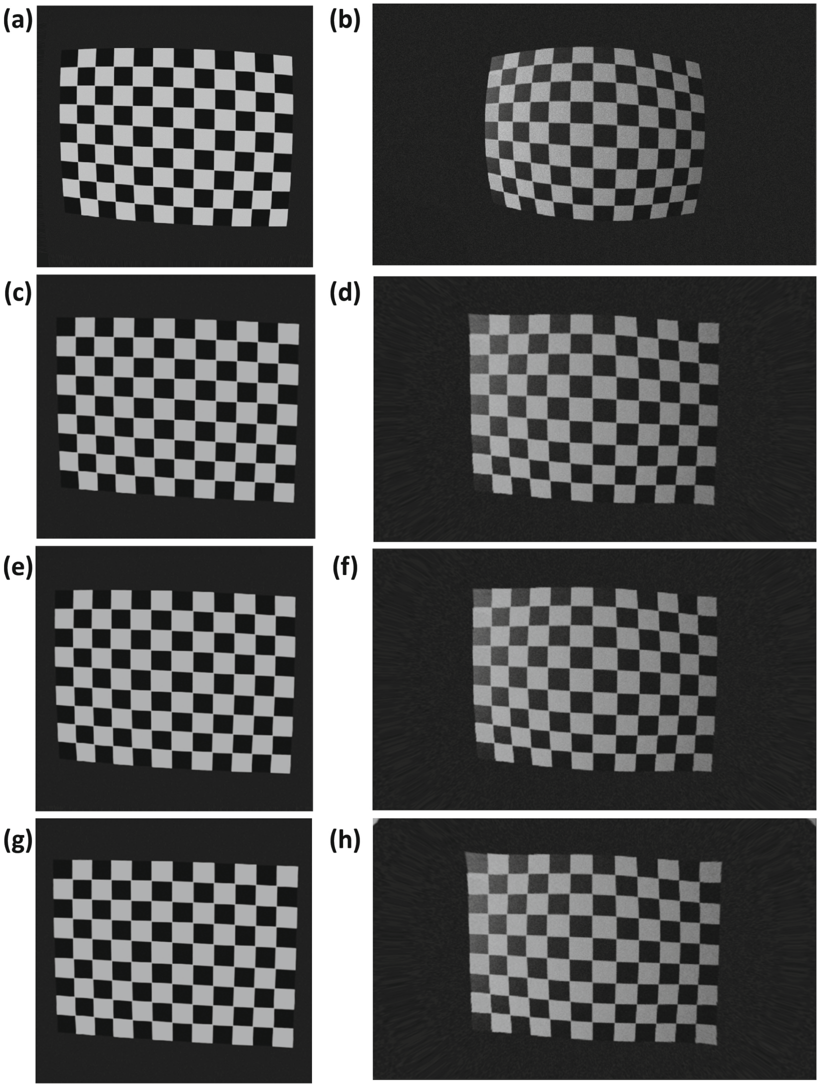


Fig. 3. (a, b) Original clean and noisy capture, undistorted render using the camera parameters predicted by (c, d) Zhang, (e, f) Tsai, (g, h) Bouguet method.

where x_d and y_d are the distorted image coordinates, the distortion model d maps the distorted image coordinates to the corrected image coordinates. The captures are saved in PNG formats, while camera parameters are saved as NumPy arrays.

For a multicamera system, the synthetic dataset generation process involves considering the intrinsic matrices of multiple cameras and simulating their interactions. In this scenario, let's denote the intrinsic matrices of the two cameras as K_1 and K_2 respectively, with corresponding distortion parameters d_1 and d_2. The extrinsic matrices P_1 and P_2 represent the position and orientation of the cameras in the global coordinate system.

To extend the methodology for a multicamera setup, the calibration pattern is moved along a trajectory that ensures visibility from both cameras. The conical spring-like path is designed to capture the pattern from varying angles, sizes, and orientations for each camera while maintaining consistency. The rotation matrices R_1 and R_2, as well as translation vectors t_1 and t_2, are determined individually for each camera frame.

The distortion parameters d_1 and d_2 are applied separately to the distorted image coordinates x_{d1} and x_{d2} of each camera. The final mapping equation for a point X in the global coordinate system to its respective distorted image coordinates x_1 and x_2 in the image planes of Camera 1 and Camera 2 with scale factor s_i is given by:

$$\begin{bmatrix} s_1 x_1 \\ s_2 x_2 \end{bmatrix} = \begin{bmatrix} K_1[R_1 \mid t_1] \\ K_2[R_2 \mid t_2] \end{bmatrix} X + \begin{bmatrix} d_1\left(\frac{x_{d1}}{f_{x1}}, \frac{y_{d1}}{f_{y1}}\right) \\ d_2\left(\frac{x_{d2}}{f_{x2}}, \frac{y_{d2}}{f_{y2}}\right) \end{bmatrix} \tag{5}$$

In this formulation, the intrinsic matrices K_1 and K_2 encapsulate the parameters specific to each camera, while the distortion parameters d_1 and d_2 account for individual radial and tangential distortions. The resulting synthetic dataset includes images from both cameras, with their respective intrinsic and extrinsic parameters saved for each frame in the dataset.

3 Results

3.1 Dataset

We created a dataset of four widely used distinct pattern types that are a 9×12 checkerboard pattern with a checker width of 15 mm, one 10×10 symmetric circle pattern with a 7 mm circle diameter, and 15 mm circle spacing, one 9×10 asymmetric circle pattern with 9 mm diameter, and 22 mm diagonal spacing and 9×12 Charuco pattern checker width of 15 mm and ArUco dictionary [18] of 7×7. Two distinct camera configurations representing a high-resolution rectilinear lens with focal length (3000, 3000), principal point (2048, 1536) with distortion parameters $[0.05, 0.02, 0.001, 0, 0, 0, 0, 0]$ and a low resolution wide, angle lens with focal length (600, 450), principal point (320, 240) with distortion parameters $[0.5, 0.1, 0.03, 0, 0, 0, 0, 0]$ are simulated for capturing the patterns. Multiple cameras with either camera configuration are added to the scene to create a stereo dataset. Skew and tangential distortion are kept at zero for all camera configurations. The extrinsic parameters R a 3×3 identity matrix and t a 3×1 zero vector are calculated using vector calculation with the relative position and orientation of the camera and target pattern to establish the ground truth. Two different external lighting conditions are used while rendering, one with uniform light across the scene without noise (Clean) and another with Directional lights with additive Gaussian noise (Noisy) in the camera captures. We created 40 data configurations with 127 captures with camera intrinsic and extrinsic matrix for each configuration with mono and stereo settings as specified in Table 1.

Table 1. Dataset statistics specifying eight different configurations based on two camera types, four pattern types, and two different environment factors available in the dataset.

Camera	Pattern	Environment
Rectilinear lens	9×12 Ch, 10×10 Sc,	Clean
(mono + stereo)	9×10 Ac, 9×12	Noisy
Wide angle lens	9×12 Ch, 10×10 Sc,	Clean
(mono + stereo)	9×10 Ac, 9×12 Cu	Noisy
Rectilinear + Wide angle	9×12 Ch, 10×10 Sc,	Clean
(stereo)	9×10 Ac, 9×12 Cu	Noisy

Table 2. RPE_{RMS} and RMSE calculated for different lens using different camera calibration methods for 9×12 Ch.

Camera	Algorithm	RPE_{RMS}	RMSE
Rectilinear lens.	Zhang's method	0.510	1.221
	Tsai's method	0.551	1.880
	Bouguet method	**0.373**	**1.127**
Wide angle lens.	Zhang's method	1.316	2.219
	Tsai's method	1.433	2.344
	Bouguet method	**0.811**	**1.861**

3.2 Evaluation

We used RMS Reprojection Error (RPE_{RMS}) as a metric to compare the algorithms and calibration patterns which can be defined as:

$$\text{RPE}_{\text{RMS}} = \sqrt{\frac{1}{N}\sum_{i=1}^{N} \|\mathbf{x}_i - \hat{\mathbf{x}}_i\|^2} \tag{6}$$

Table 3. RPE_{RMS} and RMSE calculated for four different calibration patterns in two different environmental conditions.

Environment	Pattern	Zhang's Method		Tsai's Method		Bouguet's Method	
		RPE_{RMS}	RMSE	RPE_{RMS}	RMSE	RPE_{RMS}	RMSE
Clean	9×12 Ch	0.510	1.221	0.561	1.233	**0.483**	**1.166**
	10×10 Sc	0.508	1.206	0.559	1.216	**0.490**	**1.214**
	9×10 Ac	0.506	1.205	0.554	1.198	**0.497**	**1.193**
	9×12 Cu	0.493	1.093	0.542	1.139	**0.476**	**1.012**
Noisy	9×12 Ch	1.116	**2.219**	1.227	2.290	**1.062**	2.252
	10×10 Sc	1.20	**2.263**	1.320	2.347	**1.140**	2.373
	9×10 Ac	1.118	**2.261**	1.234	2.224	**1.155**	2.319
	9×12 Cu	0.898	**1.916**	0.987	1.868	**0.853**	2.020

Table 4. RMSE_{cal} calculated over the global position of the calibration pattern for different multi-camera systems for 9×12 Ch using the triangulation method.

		Zhang's Method	Tsai's Method	Bouguet's Method
Cameras	Environment	RMSE_{cal}		
2×Rectilinear lens	Clean	2.365	2.456	**2.143**
	Noisy	**4.604**	4.812	4.746
2×Wide angle lens	Clean	3.780	3.998	**3.612**
	Noisy	**5.611**	5.742	5.798
Rectilinear & wide angle lens	Clean	4.118	4.236	**4.052**
	Noisy	**6.401**	6.552	6.577

where N is the number of points, $\mathbf{x}_i$ is the observed image point in the captured image, and $\hat{\mathbf{x}}_i$ is the corresponding projected image point using the estimated intrinsic and extrinsic parameters from the camera calibration. We also calculated accuracy by comparing the estimated intrinsic and extrinsic parameters of the camera to the ground truth values using Root Mean Square Error (RMSE) that can be defined as:

$$\text{RMSE} = \sqrt{\frac{1}{L}\sum_{i=1}^{L}\left(X_i - \hat{X}_i\right)^2} \tag{7}$$

where L is the number of parameters being estimated, X_i is the ground truth value for the i-th parameter, and $\hat{X}_i$ is the estimated value for the i-th parameter.

In our multicamera setup, we employed stereo triangulation [9] using the following equation to calculate the global position (X) of a calibration pattern, assuming knowledge of the camera parameters for two cameras within the system:

$$X = \frac{(X_1 - t_1) \times (X_2 - t_2)}{\|(X_1 - t_1) \times (X_2 - t_2)\|} \tag{8}$$

Here, X_1 and X_2 represent the 3D points in the coordinate systems of Camera 1 and Camera 2, respectively, and t_1 and t_2 are the translation vectors of Camera 1 and Camera 2. Subsequently, we utilized the Root Mean Square Error (RMSE) to quantify the disparity between the calculated global positions and the corresponding positions extracted from a simulation. The RMSE equation for a set of XYZ points involves calculating the square root of the average of the squared differences between the simulated XYZ coordinates (X_{sim}) and the calculated XYZ coordinates (X_{calc}):

$$\text{RMSE}_{\text{cal}} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}((X_{sim_i} - X_{calc_i})^2)} \tag{9}$$

This facilitated a comprehensive evaluation of the accuracy of our global position calculations in comparison to simulated ground truth, providing a quantitative measure through the $RMSE_{cal}$ metric (which is impossible to estimate in a real-world).

3.3 Analysis

Validation of SynthCal. We conducted an extensive assessment of three distinct camera calibration algorithms, taking into account both rectilinear and wide-angle camera configurations. The dataset employed in this study was generated using a 9×12 checkerboard pattern, and the findings are summarized in Table 2. Our analysis indicates that Bouget's method outperforms both Zhang's and Tsai's methods. This outcome is consistent with previous research, specifically validating the established efficacy of various calibration methods reported by Zollner et al. [23]. Upon examining the table, it becomes apparent that calibration accuracy decreases in the case of wide-angle lenses depicted in Fig. 3. This observation aligns with expectations, considering the inherent complexities associated with wide-angle lenses compared to rectilinear lenses. This discovery further underscores the reliability of our synthetic benchmark, demonstrating that center-based patterns are more effective than edge-based patterns in challenging environments. However, the increased complexity of wide-angle lenses negatively impacts their performance compared to edge-based patterns. The Charuco pattern, achieving the highest score, demonstrates its robustness to noise compared to other patterns, indicating its alignment with real-world data.

Monocular Configuration. In the context of monocular settings, our objective was to assess the effectiveness of various camera calibration patterns with different calibration algorithms. We conducted a comprehensive analysis by calculating both the RPE_{RMS} and RMSE for all eight available configurations using

Zhang's method, as detailed in Table 3. Our observations revealed that under normal conditions, the Charuco pattern consistently yielded the best results across various camera calibration algorithms. Interestingly, Bouguet's method consistently exhibited the least amount of error, regardless of the pattern type. The combination of both methods produced optimal results in terms of RPE_{RMS} and RMSE metrics. In the presence of noise, our findings indicated that irrespective of the calibration pattern type, Zhang's method outperformed Bouguet's method. This distinction was particularly evident when considering RMSE metrics, as opposed to the traditional RPE_{RMS} metrics. We attribute this phenomenon to algorithmic differences and the robustness of calibration algorithms to noise. These insights highlight the importance of considering RMSE metrics when evaluating the performance of camera calibration patterns and algorithms in monocular settings.

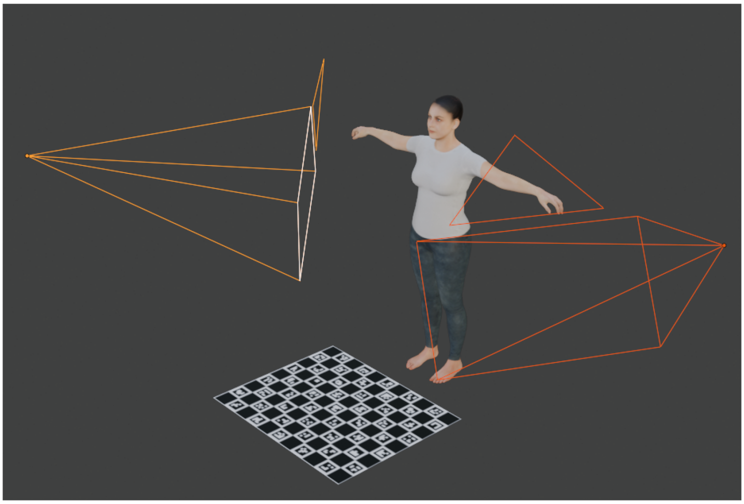


Fig. 4. Using SynthCal pipeline with DMCB [15] and EasyMocap [7] to estimate 3D pose from multiple view points

Multi-camera Configuration. Due to the availability of the absolute position of the calibration pattern, we employed $RMSE_{cal}$ to quantitatively assess the accuracy of various camera setups and calibration algorithms in multicamera settings, as outlined in Table 4. Across all camera setups in clean environments, Bouget's method consistently outperformed both Zhang's and Tsai's methods. However, in noisy setups, Zhang's method exhibited greater accuracy compared to Bouget's method. Interestingly, in some camera setups involving wide-angle lenses, Bouget's method performed worse than Tsai's method, reaffirming our earlier observation of Bouget's method's lack of robustness in noisy conditions.

Table 5. MPJPE calculated by comparing the original 24 joint SMPL pose with the predicted SMPL pose using EasyMocap [7] for different calibration configurations.

		Zhang's Method	Tsai's Method	Bouguet's Method
Cameras	Environment	MPJPE		
2×Rectilinear	Clean	5.974	6.011	**5.967**
lens.	Noisy	**6.313**	6.381	6.344
2×Wide angle	Clean	5.986	6.028	**5.962**
lens.	Noisy	**6.440**	6.485	6.478
Rectilinear & wide	Clean	6.212	6.343	**6.413**
angle lens	Noisy	**6.511**	6.595	6.620

Additionally, our observations revealed that, for stereo setups, identical camera pairs demonstrated superior performance compared to non-identical pairs. This trend persisted even when wide-angle lenses, known for their complexity, were involved. Surprisingly, the combination of two rectilinear lenses consistently outperformed setups comprising one wide-angle and one rectilinear lens.

To further validate our model and assess the accuracy of 3D pose estimation across various camera configurations and calibration algorithms, we utilized the DMCB [15] to simulate a textured human mesh in SMPL [14] format using the motion imported from TotalCapture dataset [16] and texture imported from SMPLitex [4] within Blender as visualized in Fig. 4. This simulated mesh was then captured by multiple cameras positioned at different angles, employing different calibration algorithms and calibration patterns using SynthCal. The rendered videos are then given as input to EasyMocap [7] to estimate the 24-joint SMPL pose.

Unlike real-world scenarios where ground truth pose data might be unavailable, here we have access to the real ground truth of the 3D pose, as it was used to create the animated mesh. By comparing the estimated 3D poses with this ground truth, measured through metrics like Mean Per Joint Position Error (MPJPE) which is defined as:

$$MPJPE = \frac{1}{N} \sum_{i=1}^{N} \sqrt{\sum_{j=1}^{J} \|\mathbf{P}_{ij} - \mathbf{P}_{ij}^{GT}\|_2^2}$$

where N is the number of frames, J is the number of joints, $\mathbf{P}_{ij}$ denotes the estimated joint position, and $\mathbf{P}_{ij}^{GT}$ represents the ground truth joint position, we were able to validate our findings.

The results given in Table 5 follow a similar trend already established in Table 4 hence validating our benchmark. Our results underscored the impact of calibration patterns and algorithms on the accuracy of 3D pose estimation models, although not to a very high extent. This validation highlights the importance of meticulous calibration procedures and algorithm selection in enhancing the accuracy and reliability of machine learning models for tasks like 3D pose estimation.

4 Conclusion

In this paper, we introduced the SynthCal pipeline evaluating camera calibration methods. Our research underscores the efficacy of the Charuco pattern coupled with Bouguet's method under standard conditions, while Zhang's method demonstrates superiority in noisy environments. Bouguet's approach fares admirably in pristine setups but encounters challenges with wide-angle lenses and noise. Consistency among camera pairs surpasses mixed configurations, underscoring the significance of uniformity. Our results demonstrated the importance of considering diverse metrics in calibration assessment.

For future work, we could expand SynthCal to incorporate non-planar calibration algorithms, thereby enhancing its relevance across various camera models and applications.

Acknowledgements. The research reported in this paper was supported by the BMBF in the project VidGenSense (01IW21003).

References

1. An, G.H., Lee, S., Seo, M.W., Yun, K., Cheong, W.S., Kang, S.J.: Charuco board-based omnidirectional camera calibration method. Electronics **7**, 421 (2018)
2. Barbero-García, I., Lerma, J.L., Miranda, P., Marqués-Mateu, Á.: Smartphone-based photogrammetric 3D modelling assessment by comparison with radiological medical imaging for cranial deformation analysis. Measurement **131**, 372–379 (2019)
3. Bouguet, J.Y.: Camera calibration toolbox for MatLAB (2004). http://www.vision.caltech.edu/bouguetj/calib_doc/index.html
4. Casas, D., Trinidad, M.C.: SMPLITEX: a generative model and dataset for 3D human texture estimation from single image. arXiv preprint (2023)
5. Chen, J., Little, J.J.: Sports camera calibration via synthetic data. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (2019)
6. Chuang, J.H., Ho, C.H., Umam, A., Chen, H.Y., Hwang, J.N., Chen, T.A.: Geometry-based camera calibration using closed-form solution of principal line. IEEE Trans. Image Process. **30**, 2599–2610 (2021)
7. Dong, J., Fang, Q., Jiang, W., Yang, Y., Bao, H., Zhou, X.: Fast and robust multi-person 3D pose estimation and tracking from multiple views. In: T-PAMI (2021)
8. Feng, D., Rosenbaum, L., Glaeser, C., Timm, F., Dietmayer, K.: Can we trust you? On calibration of a probabilistic object detector for autonomous driving. arXiv preprint (2019)
9. Hahne, C., Aggoun, A., Velisavljevic, V., Fiebig, S., Pesch, M.: Baseline and triangulation geometry in a standard plenoptic camera. Int. J. Comput. Vis. **126**, 21–35 (2018)

10. Honauer, K., Johannsen, O., Kondermann, D., Goldluecke, B.: A dataset and evaluation methodology for depth estimation on 4D light fields. In: 13th Asian Conference on Computer Vision (2017)
11. Huai, J., et al.: A review and comparative study of close-range geometric camera calibration tools. arXiv preprint (2023)
12. Kang, Z., Yang, J., Yang, Z., Cheng, S.: A review of techniques for 3D reconstruction of indoor environments. ISPRS Int. J. Geo-Inform. **9**, 330 (2020)
13. Kikkawa, S., Okura, F., Muramatsu, D., Yagi, Y., Saito, H.: Accuracy evaluation and prediction of single-image camera calibration. IEEE Access **11**, 19312–19323 (2023)
14. Loper, M., Mahmood, N., Romero, J., Pons-Moll, G., Black, M.J.: SMPL: a skinned multi-person linear model. In: Seminal Graphics Papers: Pushing the Boundaries, vol. 2, pp. 851–866 (2023)
15. Ray, L.S.S., Zhou, B., Suh, S., Lukowicz, P.: Selecting the motion ground truth for loose-fitting wearables: Benchmarking optical mocap methods. In: the 2023 ACM International Symposium on Wearable Computers (2023)
16. Trumble, M., Gilbert, A., Malleson, C., Hilton, A., Collomosse, J.: Total capture: 3D human pose estimation fusing video and inertial sensors. In: Proceedings of 28th British Machine Vision Conference (2017)
17. Tsai, R.: A versatile camera calibration technique for high-accuracy 3D machine vision metrology using off-the-shelf tv cameras and lenses. IEEE J. Robot. Autom. **3**, 323–344 (1987)
18. Tzortzis, G., Likas, A.: The minmax k-means clustering algorithm. Pattern Recogn. **47**, 2505–2516 (2014)
19. Usamentiaga, R., Ibarra-Castanedo, C., Maldague, X.: Comparison and evaluation of geometric calibration methods for infrared cameras to perform metric measurements on a plane. Appl. Opt. **57**, D1–D10 (2018)
20. Wu, C.Y., Johnson, J., Malik, J., Feichtenhofer, C., Gkioxari, G.: Multiview compressive coding for 3D reconstruction. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2023)
21. Zhang, Y., Wang, T., Zhang, X.: MOTRV2: nootstrapping end-to-end multi-object tracking by pretrained object detectors. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2023)
22. Zhang, Z.: A flexible new technique for camera calibration. IEEE Trans. Pattern Anal. Mach. Intell. **22**, 1330–1334 (2000)
23. Zollner, H., Sablatnig, R.: Comparison of methods for geometric camera calibration using planar calibration targets (2004)

Arbitrary Clothing Style Transfer Based on Attention Mechanism

Chen Huang[1](✉), Junjie Zhang[1,2], and Hua Yuan[3]

[1] School of Computer Science and Artificial Intelligence, Wuhan Textile University, Wuhan 430200, China
2215063031@mail.wtu.edu.cn, 2007086@wtu.edu.cn

[2] Hubei Key Laboratory of Digital Textile Equipment, Wuhan Textile University, Wuhan 430073, China

[3] Wuhan Textile and Apparel Digital Engineering Technology Research Center, School of Fashion, Wuhan Textile University, Wuhan 430073, China
2019009@wtu.edu.cn

Abstract. In recent years, style transfer has become increasingly prominent in various domains, especially fashion. As a tool for designers, clothing style transfer generates a wide array of styles, enabling rapid experimentation and fostering creative inspiration. However, current methods have poor performance in transferring textures and colors from style images to clothing, and commonly result in blurred boundaries between clothing and background. To address these challenges, an arbitrary style transfer algorithm for clothing is proposed, utilizing attention network and feature fusion for more effective and efficient style application. In this paper, the criss-cross attention network is incorporated to extract comprehensively global features and capture long-range dependencies, thus minimizing artifacts and enhancing texture transfer. Through a novel multi-level feature fusion approach, color transfer becomes more natural and coherent, closely aligning with the color palette of the style image. Additionally, semantic segmentation is employed to separate clothing from the background, preserving the original background and character. The experimental results show that the user preference of this paper's algorithm is much higher than existing methods, and single 512×512 image style transfer takes only 0.314 s with real-time performance.

Keywords: Style transfer · Criss-cross attention · Feature fusion · Semantic segmentation

1 Introduction

As the overall economic level of society improves, there is a growing demand for personalized clothing styles. Traditionally, clothing design relies heavily on the designer's experience, requiring significant time for conceptualization and numerous sketches to compare different styles. This process is labor-intensive and resource-consuming. Clothing style transfer addresses these challenges by

A. Antonacopoulos et al. (Eds.): ICPR 2024, LNCS 15332, pp. 239–252, 2025.
https://doi.org/10.1007/978-3-031-78125-4_17

enabling the quick and easy application of various styles to clothing, allowing designers to rapidly experiment with and compare designs. This technology not only provides abundant design inspiration but also shortens the design cycle and improves design efficiency.

Several methods for style transfer have been developed, including the pioneering work by Gatys et al. [3], adaptive instance normalization (AdaIN) [5], style attention network (SANet) [12], multi-adaption style transfer(MAST) [2], internal-external learning and contrastive learning(IECAST) [1], exact feature distribution matching(EFDM) [22], and content affinity preserved versatile style transfer(CAP-VSTNet) [20]. Although capable of single style, multi-style, and arbitrary style transfer, these methods often struggle to effectively transfer texture and color from the style image to the clothing image. This results in local artifacts and diminishes visual quality in the stylized images. Thus, there remains a need for improved methods to enhance the visual effects and accuracy of style transfer in clothing images.

Moreover, due to the broad application prospects of style transfer in clothing, researchers have increasingly focused on exploring methods for conducting style transfer specifically in the realm of fashion. Sbai et al. [15] introduced a StyleGAN-based algorithm for generating clothing design images, while Liu et al. [9] proposed an Attribute-GAN model to tackle clothing matching issues. Han et al. [4] developed fashion inpainting networks to synthesize diverse and compatible clothing images, and Yu et al. [21] designed a network structure to match user preferences and clothing design creation. Despite their advancements, these methods primarily execute global style transfer on clothing images, potentially distorting crucial background features such as facial characteristics and yielding subpar transfer results. To address this, researchers have proposed methods that preserve the background during style transfer. Mo et al. [11] enhanced InstaGAN using a CycleGAN-based approach, incorporating semantic segmentation to separate original image information before transferring styles. Similarly, Sun et al. [16] proposed a two-stage unsupervised approach, constructing a network to unsupervisedly split out the clothing texture region in the first stage.

To better assist clothing designers, clothing style transfer should avoid spatial distortion and preserve meaningful information in the clothing image, ensuring that the color and texture of the stylized image are harmonious and coherent. Moreover, preserving important features such as faces in the background of the clothing image is crucial. To achieve these objectives, this paper proposes an arbitrary style transfer algorithm for clothing based on attention mechanism. This algorithm can synthesize high-quality stylized images while retaining the original background and character. The main contributions of this paper can be summarized as follows:

- This paper proposes an arbitrary clothing style transfer network that incorporates the criss-cross attention mechanism. This network comprehensively extracts the global features of the clothing image and captures long-range dependencies, reducing artifacts in the stylized image and enhancing texture transfer.

- This paper proposes the Criss-cross Attention Feature Fusion Module (CCAFFM). This module employs an innovative multi-level feature fusion method, facilitating enhanced transfer of the style image's color to the clothing, resulting in high-quality stylized clothing images with natural and harmonious color tones.
- This paper incorporates a semantic segmentation network to separate clothing from the background. This enables the creation of stylized clothing images that retain the original background and character.

2 Related Work

2.1 Arbitrary Style Transfer

Style transfer, originating from non-photorealistic rendering, is closely associated with texture synthesis and texture transfer. Gatys et al. [3] pioneered a method using Convolutional Neural Network (CNN) to achieve style transfer by minimizing the difference in feature representations between content and style images. [7,8,17] proposed real-time feed-forward style transfer networks for real-time applications, although these typically require separate training for each style. To broaden the applicability of style transfer, arbitrary style transfer has emerged as a key research area, with significant efforts dedicated to improving its efficiency and effectiveness.

The AdaIN algorithm [5] adjusts the mean and variance of content images to match those of style images by utilizing global feature statistics, effectively transferring texture but exhibiting limitations in color transfer. To overcome the shortcomings of AdaIN, the SANet [12] incorporates a self-attention-based module for arbitrary style transfer, though it can result in the loss of crucial information and artifacts in localized regions. The MAST algorithm [2] introduces adaptive modules for capturing a wide range of styles but incurs significant computational costs. The IECAST algorithm [1] employs contrastive losses to facilitate simultaneous learning from individual style images and large-scale style datasets, yet it may not effectively transfer the textures and colors of the target style. The EFDM [22] performs cross-distribution feature matching in a single step, providing a more precise measurement of distribution divergence. Although its texture transfer effect improves, the color transfer is still insufficient. The CAP-VSTNet [20] uses a reversible residual network to preserve content affinity and reduce redundant information. Despite the improvements, it still fails to effectively address the issue of color transfer, leaving much to be desired in this aspect.

The existing arbitrary style transfer methods often struggle to effectively coordinate local and global styles, transfer texture and color from the style image, and generate high-quality stylized images. To address these limitations, this paper proposes an advanced arbitrary style transfer algorithm known as the Criss-cross Attention Feature Fusion Module (CCAFFM). This algorithm introduces criss-cross attention network to comprehensively extract global features and capture long-range dependency relationships, allowing for the flexible matching of style features based on the semantic spatial distribution of clothing images. Additionally, this paper presents a novel multi-level feature fusion method that improves the transfer of color from the style image to the clothing. The proposed module effectively reduces artifacts in the stylized images and significantly improves the transfer of texture and color, resulting in high-quality stylized outputs.

2.2 Attention

In contemporary deep learning tasks, attention mechanisms have proven to be highly effective. Unlike traditional models that compress entire images into static representations, attention mechanisms enable models to focus dynamically on the most relevant parts and features of an image. Vaswani et al. [18] introduced the self-attention mechanism, enabling models to establish associations between different locations within the same sequence. This mechanism computes the output at each position as a weighted sum of all positions in the input sequence, with weights determined by an attention distribution. To address the limitations of receptive fields, Wang et al. [19] proposed a non-local attention mechanism that captures global information rather than just local areas. This non-local approach improves feature representation by considering relationships across the entire sequence, but it is significantly constrained by computational complexity. In order to reduce the amount of computation, Huang et al. [6] proposed the criss-cross attention mechanism, which reduces computational load while effectively capturing and fusing context information along criss-cross paths. This approach enhances the model's semantic understanding capabilities and addresses the issue of computational complexity.

3 Approach

3.1 Network Architecture

The proposed network takes a clothing image I_c and a style image I_s to synthesize a clothing stylized image I_{cs}. In the proposed network, a pre-trained VGG-19 network is employed as the encoder to extract multi-scale feature maps. This encoder consists of a series of convolutional and pooling layers. The decoder uses a symmetric structure of VGG-19. As shown in Fig. 1, firstly, the VGG feature maps F_c and F_s are extracted at each layer of the encoder from a clothing image I_c and style image I_s pair, including $Relu_{1\_1}$, $Relu_{2\_1}$, $Relu_{3\_1}$, $Relu_{4\_1}$ and $Relu_{5\_1}$. After encoding the clothing and style images, the feature maps from $Relu_{4\_1}$ and $Relu_{5\_1}$ are fed into CCAFFM module. This module maps the correspondences between the clothing features and the style features, producing the stylized feature map F_{csc}:

$$F_{csc} = CCAFFM\left(F_{c\_4\_1}, F_{c\_5\_1}, F_{s\_4\_1}, F_{s\_5\_1}\right). \tag{1}$$

Then, a 3×3 convolution operation is applied to produce the final stylized feature map. This map is then fed into the decoder to reconstruct the global

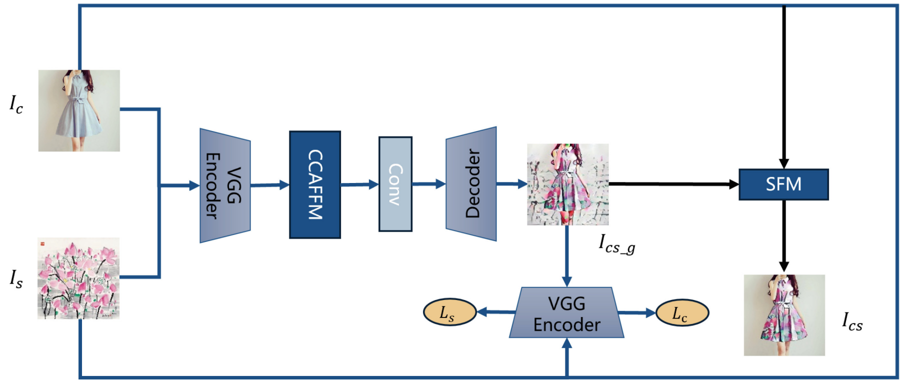
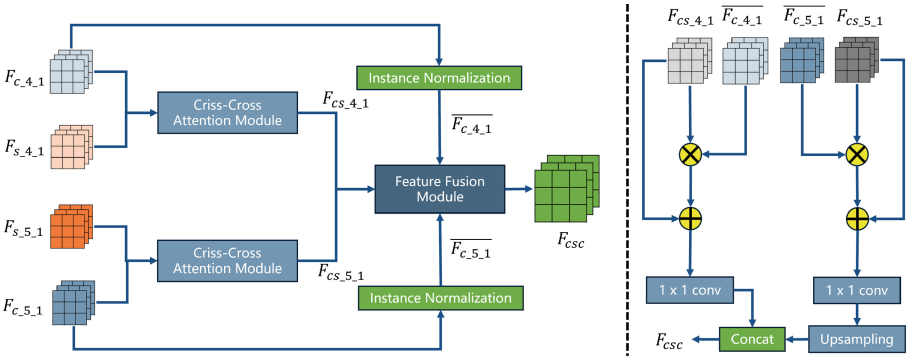


Fig. 1. Structure of proposed arbitrary clothing style transfer network.

stylized image. Finally, I_c and I_{cs_g} are passed through the Segmentation Fusion Module(SFM), yielding the final clothing-stylized image I_{cs}.



(*a*) *The CCAFFM*

(*b*) *The Feature Fusion Module*

Fig. 2. The overall structure of the Criss-cross Attention Feature Fusion Module (CCAFFM).

3.2 Criss-Cross Attention Feature Fusion Module (CCAFFM)

The Criss-cross Attention Feature Fusion Module (CCAFFM) proposed in this paper is shown in Fig. 2. Firstly, the clothing feature maps, $F_{c\_4\_1}$ and $F_{c\_5\_1}$, and the style feature maps, $F_{s\_4\_1}$ and $F_{s\_5\_1}$, extracted from the encoder are input into the criss-cross attention module to conduct spatial reorganization, resulting in the feature maps $F_{cs\_4\_1}$ and $F_{cs\_5\_1}$. In addition, the clothing feature maps, $F_{c\_4\_1}$ and $F_{c\_5\_1}$, separately undergo instance normalization to get $\overline{F_{c\_4\_1}}$ and

$\overline{F_{c\_5\_1}}$. The formula is as follows:

$$\overline{F_c} = \gamma\left(\frac{F_c - \mu\left(F_c\right)}{\sigma\left(F_c\right)}\right) + \beta, \tag{2}$$

where γ and β are parameters learned from data. Then, $F_{cs\_4\_1}$, $F_{cs\_5\_1}$, $\overline{F_{c\_4\_1}}$ and $\overline{F_{c\_5\_1}}$ are input into the feature fusion module to get the stylized feature map F_{csc}.

During the feature fusion process, as shown in Fig. 2, firstly, the feature maps $F_{cs\_4\_1}$ and $F_{cs\_5\_1}$ are initially multiplied by their corresponding normalized clothing feature maps $\overline{F_{c\_4\_1}}$ and $\overline{F_{c\_5\_1}}$. The resulting products are then summed with the original feature maps $F_{cs\_4\_1}$ and $F_{cs\_5\_1}$, respectively. Following this step, a 1×1 convolution and an upsampling operation are applied. Finally, the results are concatenated to generate the stylized feature map F_{csc}. The process can be summarized by the following formula:

$$\begin{aligned} F_{csc} = & concat(conv(F_{cs\_4\_1} * \overline{F_{c\_4\_1}} + F_{cs\_4\_1}), \\ & upsampling(conv(F_{cs\_5\_1} * \overline{F_{c\_5\_1}} + F_{cs\_5\_1}))). \end{aligned} \tag{3}$$

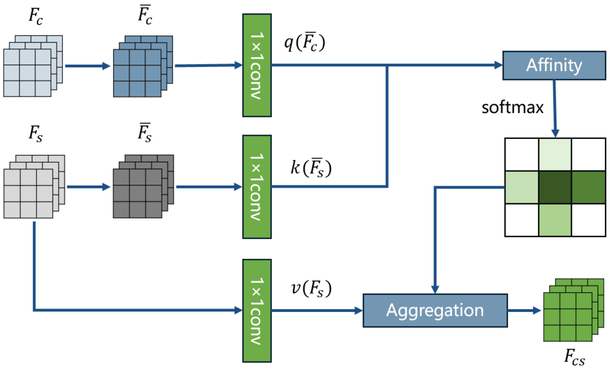


Fig. 3. The Criss-cross Attention Mechanism.

As illustrated in Fig. 3, the criss-cross attention network processes the clothing feature map F_c and the style feature map F_s by first normalizing them to $\overline{F}_c$ and $\overline{F}_s$, respectively. Following normalization, $\overline{F}_c$, $\overline{F}_s$ and F_s each pass through a 1×1 convolutional layer, generating feature maps $q(\overline{F}_c)$, $k(\overline{F}_s)$ and $v(F_s)$. Here, $v(F_s)$ represents the output response of the style feature at position j. The Affinity operation then computes the similarity between the position i of clothing feature and all positions j of style feature within the same row and column as position i. This operation is defined as follows:

$$f(\overline{F}_c, \overline{F}_s) = q(\overline{F}_c)^{\mathrm{T}} k(\overline{F}_s). \tag{4}$$

The results are then converted into probability distributions by softmax operation to better represent the relative importance between different positions, so that the model can better capture and understand the relationship between clothing features and style features. Aggregation operation fuses the original style feature map with the feature map output by softmax to obtain the clothing style feature map F_{cs}.

The incorporation of the criss-cross attention network in the proposed Criss-cross Attention Feature Fusion Module (CCAFFM) enables the precise mapping of relationships between clothing features and style features. This approach embeds local style features into the appropriate positions within the clothing features and integrates global style patterns efficiently and flexibly. Moreover, by utilizing a novel multi-level feature fusion method, this paper's approach enhances significant features while reducing the impact of less important ones, thereby improving the accuracy of color transfer from style images. Consequently, the CCAFFM synthesizes high-quality stylized images in real time and markedly enhances the transfer of textures and colors to the clothing image.

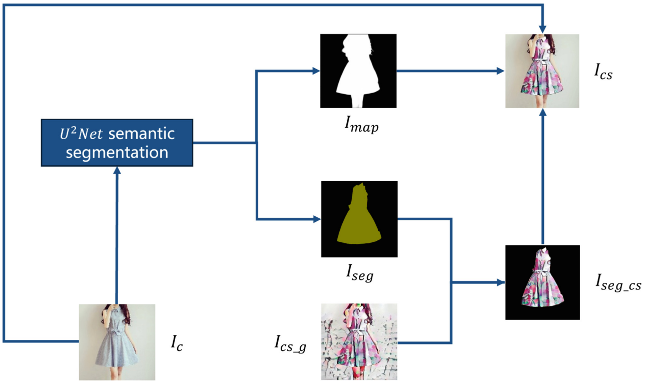


Fig. 4. The Segmentation Fusion Module (SFM).

3.3 Segmentation Fusion Module (SFM)

To retain the original background and character, this paper presents the Segmentation Fusion Module (SFM). As shown in Fig. 4, using a pre-trained U^2Net model [14], semantic segmentation is performed on the clothing image I_c, generating the saliency map I_{map} and the segmented image I_{seg}. The segmented image is then combined with the stylized global image I_{cs_g} to produce the stylized segmented image I_{seg_cs}. Finally, the stylized segmented image is fused with the original clothing image to create the final clothing stylized image I_{cs}. During this blending process, smooth edge processing is applied to ensure seamless integration. The saliency map I_{map} values range from 0 to 1, and the process is mathematically described as follows:

$$I_{cs} = I_{map} I_{seg\_cs} + (1 - I_{map}) I_c. \tag{5}$$

3.4 Loss Function

The overall loss function consists of two parts: the clothing loss and the style loss. As shown in Fig. 1, the VGG encoder is used to calculate the loss. The formula is as follows:

$$\mathcal{L} = \lambda_c \mathcal{L}_c + \mathcal{L}_s, \tag{6}$$

where $\mathcal{L}_c$ denotes the clothing loss, $\mathcal{L}_s$ represents the style loss. λ_c is the weight of the clothing loss.

The clothing loss is the sum of the Euclidean distances between the clothing features and the stylized features. It is defined as follows:

$$\mathcal{L}_c = \|F_{csc} - F_{c\_4\_1}\|_2 + \|F_{csc} - F_{c\_5\_1}\|_2 . \tag{7}$$

The style loss consists of two components. The first component is the sum of two Euclidean distance: one between the mean of the encoder's style features and the stylized features, and the other between the variance of the encoder's style features and the stylized features. Here, $F_{s(i)}$ represents the feature map at each layer of the encoder extracted from the style image. The second component is inspired by SANet [12] and focuses on preserving the clothing structure of the image rather than changing the style. F_{ccc} and F_{sss} represent the feature map obtained from the CCAFFM module using identical clothing images or identical style images, respectively. $F_{c(i)}$ represents the feature map at each layer of the encoder extracted from the clothing image. λ_1 and λ_2 are the weights for the style loss components. The style loss is defined as follows:

$$\begin{aligned} \mathcal{L}_s = {} & \lambda_1 \sum_{i=1}^{N} (\|\mu(F_{csc}) - \mu(F_{s(i)})\|_2 + \|\sigma(F_{csc}) - \sigma(F_{s(i)})\|_2) \\ & + \lambda_2 \sum_{i=1}^{N} \left(\|F_{ccc} - F_{c(i)}\|_2 + \|F_{sss} - F_{s(i)}\|_2\right) . \end{aligned} \tag{8}$$

The weighting parameters are set as $\lambda_c = 1$, $\lambda_1 = 3$, and $\lambda_2 = 50$ in the experiments.

4 Experimental Results

4.1 Experimental Settings

This study utilizes the DeepFashion [10] dataset for clothing images and the WikiArt [13] dataset for style images. The experiments were conducted on an

NVIDIA Tesla V100 GPU with 32 GB RAM. The Adam optimizer was employed with a learning rate of 0.0001 and a batch size of 5 for the clothing-style image pairs. Each training image was resized to 512 × 512, maintaining the aspect ratio, and then randomly cropped to 256 × 256. The training process consists of 100,000 iterations. During testing, the network shows flexibility in handling various input sizes due to its fully convolutional architecture.

4.2 Comparison with State-of-the-Art Methods

To assess the performance of the proposed method, this paper compared it with five established arbitrary style transfer techniques: AdaIN [5], SANet [12], MAST [2], IECAST [1] and EFDM [22]. This paper's approach uniquely integrates a Segmentation Fusion Module (SFM) for background separation, as illustrated in Fig. 5. This module effectively segments the clothing from the background in the generated stylized images, thereby enhancing visual quality. The semantic segmentation clarifies the differences between the stylized and original images, aiding designers in their creative processes. To ensure a fair comparison of arbitrary style transfer effectiveness, SFM is also applied to the stylized images generated by the other methods.

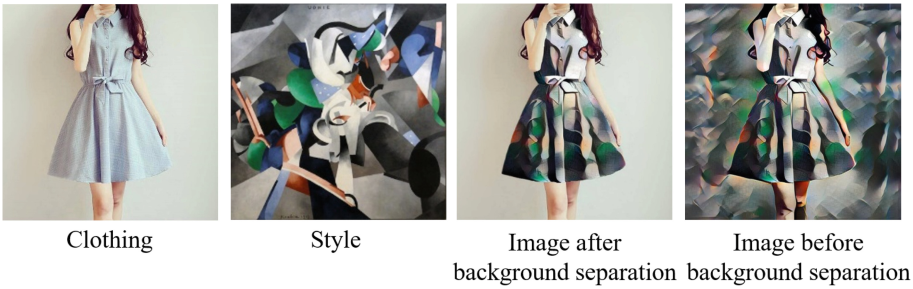

Clothing
Style
Image after background separation
Image before background separation

Fig. 5. The effect of background separation by using SFM.

Qualitative Comparison. In Fig. 6, this paper presents comparisons of clothing style transfer results among various state-of-the-art methods and the proposed approach. As shown in rows 3 in Fig. 6, AdaIN adjusts clothing feature mean and variance for stylized image generation, but its results often lack visual appeal and fail to achieve satisfactory color transfer. As shown in rows 4 in Fig. 6, SANet utilizes the weighted average sum of all pixels to represent features, yet it struggles to adjust local style features adequately, resulting in image artifacts and unclear texture structures. The MAST employs a non-local approach to calculate local similarity between clothing and style features and adjust the style feature distribution based on clothing features. However, similar to SANet method, MAST-generated stylized images suffer from artifacts and poor color transfer. As shown in rows 6 in Fig. 6, IECAST incorporates style information

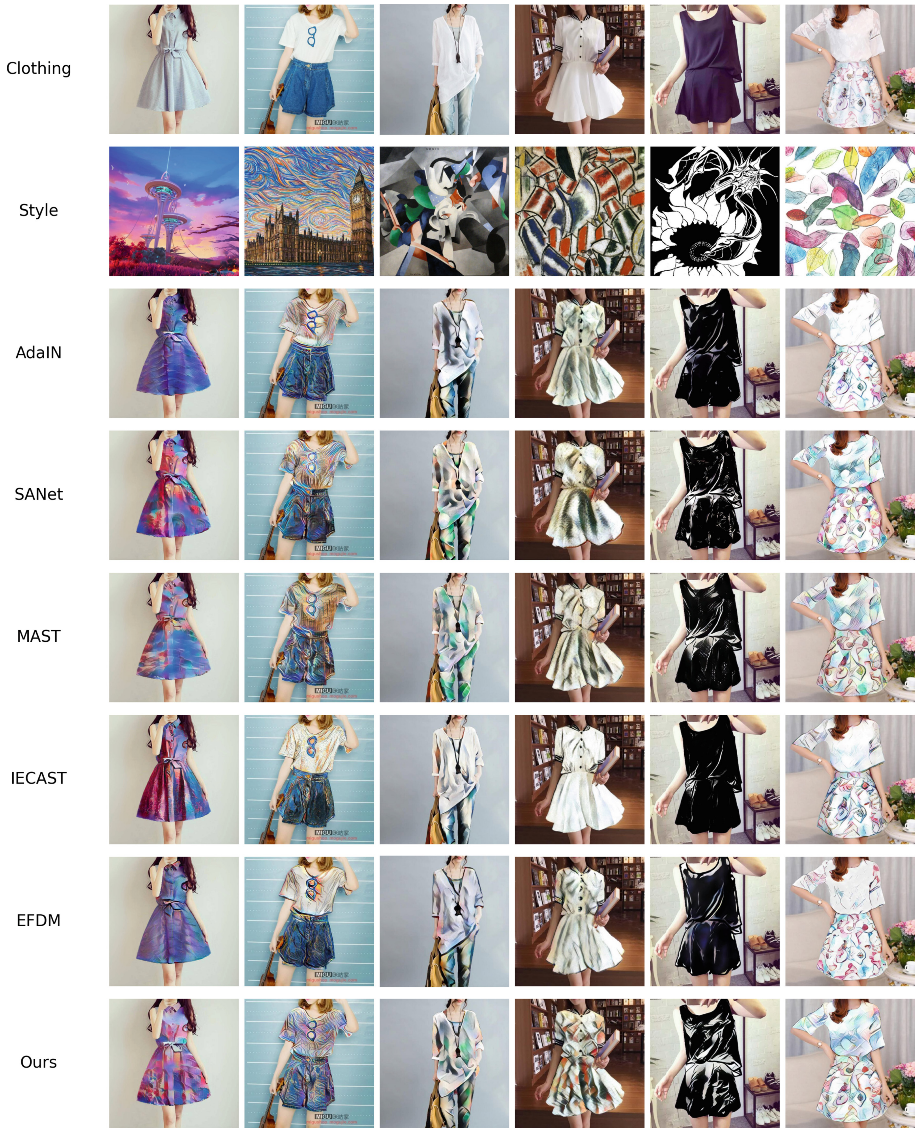


Fig. 6. Qualitative comparisons on image style transfer. The first row shows the clothing images. The second row shows the style images. The rest of the rows show the stylization results generated by different style transfer methods.

from large-scale datasets and the target style image, but it fails to effectively transfer the textures and colors of the target style image. EFDM adopts exact feature distribution matching, yet, like AdaIN, it faces challenges in color transfer, as shown in rows 7 in Fig. 6. In contrast, this paper's method demonstrates exceptional capability in generating high-fidelity stylized images characterized by clear texture definition and nuanced style representation. These images adeptly preserve the intricate texture details and vibrant color schemes inherent in the style images while faithfully capturing both global and local stylistic elements. Notably, the stylized images exhibit clearer textures and more natural color transitions compared to other methods, minimizing the occurrence of image artifacts. For example, in the eighth row and first column of Fig. 6, the stylized image demonstrates notably clear texture without variegated colors in the collar and skirt regions. Additionally, accurate color transfer of the shirt and pants in rows 8 and columns 3 of Fig. 6 contributes to a clean and crisp overall appearance with distinct textures. Furthermore, the stylized images in rows 8 and columns 4, 5 of Fig. 6 demonstrate superior style feature transfer by effectively capturing the style characteristics of different regions in the style images when compared to other methods.

User Study. Given the inherently subjective nature of artistic style transfer, this paper conducted a user study to assess the performance of the proposed method. Firstly, 10 clothing images and 50 style images were randomly selected, creating a total of 500 clothing-style pairs. From these, 20 pairs were sampled to generate stylized images using different methods. Participants were shown these images side-by-side in random order and asked to select the most appealing image based on the texture transfer effect and color transfer effect. Finally, this paper collected 800 votes from 20 users and displayed the results in a bar diagram. The results, depicted in Fig. 7, indicate that this paper's method can produce more appealing stylized images in terms of both texture and color compared to the other methods.

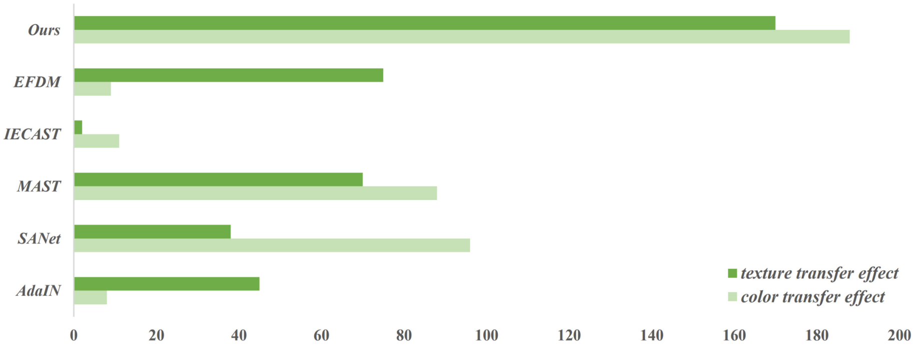


Fig. 7. User preference result of six style transfer algorithms.

Efficiency Analysis. Table 1 presents the runtime performance of the proposed method compared to other methods at image resolutions of 256 and 512 pixels. The experiments were conducted on a 4 GB NVIDIA GeForce RTX 2050 GPU. To ensure a fair evaluation, the runtime for each method was averaged over 100 images, with each image processed 10 times to mitigate variations in GPU performance. As shown in Table 1, the proposed algorithm is faster than SANet, MAST and IECAST. Despite slightly slower than AdaIN and EFDM, the difference in speed is minimal. Therefore, the proposed algorithm can be considered efficient for fast arbitrary style transfer, achieving real-time performance.

Table 1. Execution time comparison (in seconds).

Method	Time (256px)	Time (512px)
AdaIN [5]	0.065	0.245
SANet [12]	0.085	0.336
MAST [2]	0.117	0.710
IECAST [1]	0.085	0.330
EFDM [22]	0.065	0.246
Ours	0.084	0.314

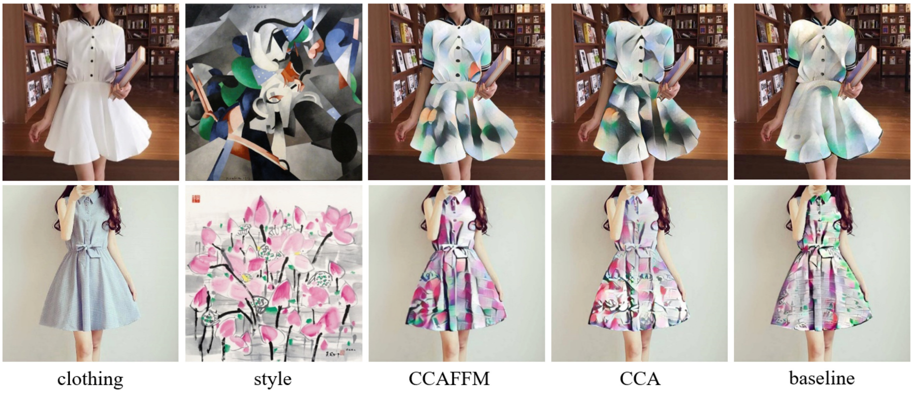


Fig. 8. Ablation studies of Criss-cross Attention Feature Fusion Module (CCAFFM).

4.3 Ablation Studies

To evaluate the effectiveness of the proposed Criss-cross Attention Feature Fusion Module (CCAFFM), ablation experiments were conducted by comparing it with two variants: utilizing only the criss-cross attention network (CCA) and the baseline model SANet. The results depicted in Fig. 8 demonstrate that compared to the baseline, integrating the CCA module enhances texture transfer and reduces artifacts in the stylized image. Furthermore, the incorporation of new feature fusion method improves the accuracy of color transfer from the style image, resulting in visually appealing outcomes. Overall, the CCAFFM module enhances texture clarity and effectively transfers both texture and color from the style image, leading to a more natural color distribution in the stylized image.

5 Conclusions

In conclusion, the proposed arbitrary clothing style transfer network in this paper presents solutions to prominent challenges in the field. Through the utilization of the criss-cross attention network, the network achieves comprehensive feature extraction and dependency relationship capture, thereby reducing artifacts and enhancing texture transfer in stylized images. Additionally, the novel multi-level feature fusion method improves color transfer from style images. By integrating a semantic segmentation network, the original background and character details are preserved in the stylized clothing images. These improvements collectively facilitate the generation of high-quality, visually appealing stylized clothing images. However, although the proposed method has achieved significant results, it needs additional semantic segmentation network to realize background separation. The future research will explore how to retain the background during the process of style transfer, ultilizing the attention mechanism to transfer style to clothing only without affecting the background.

References

1. Chen, H., et al.: Artistic style transfer with internal-external learning and contrastive learning. In: Neural Information Processing Systems (2021). https://api.semanticscholar.org/CorpusID:247546450
2. Deng, Y., Tang, F., Dong, W., Sun, W.C., Huang, F., Xu, C.: Arbitrary style transfer via multi-adaptation network. In: Proceedings of the 28th ACM International Conference on Multimedia (2020). https://api.semanticscholar.org/CorpusID:218900841
3. Gatys, L.A., Ecker, A.S., Bethge, M.: Image style transfer using convolutional neural networks. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2414–2423 (2016). https://api.semanticscholar.org/CorpusID:206593710
4. Han, X., Wu, Z., Huang, W., Scott, M.R., Davis, L.S.: Compatible and diverse fashion image inpainting. ArXiv abs/1902.01096 (2019). https://api.semanticscholar.org/CorpusID:59599920
5. Huang, X., Belongie, S.J.: Arbitrary style transfer in real-time with adaptive instance normalization. In: 2017 IEEE International Conference on Computer Vision (ICCV), pp. 1510–1519 (2017). https://api.semanticscholar.org/CorpusID:6576859
6. Huang, Z., et al.: CCNet: criss-cross attention for semantic segmentation. IEEE Trans. Pattern Anal. Mach. Intell. **45**, 6896–6908 (2018). https://api.semanticscholar.org/CorpusID:53846561
7. Johnson, J., Alahi, A., Fei-Fei, L.: Perceptual losses for real-time style transfer and super-resolution. ArXiv abs/1603.08155 (2016). https://api.semanticscholar.org/CorpusID:980236
8. Li, C., Wand, M.: Precomputed real-time texture synthesis with Markovian generative adversarial networks. In: European Conference on Computer Vision (2016). https://api.semanticscholar.org/CorpusID:18781152
9. Liu, L., Zhang, H., Ji, Y., Wu, Q.M.J., Zhang, D.Z.: Toward AI fashion design: An attribute-GAN model for clothing match. Neurocomputing **341**, 156–167 (2019). https://api.semanticscholar.org/CorpusID:126513803

10. Liu, Z., Luo, P., Qiu, S., Wang, X., Tang, X.: DeepFashion: powering robust clothes recognition and retrieval with rich annotations. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) pp. 1096–1104 (2016). https://api.semanticscholar.org/CorpusID:206593370
11. Mo, S., Cho, M., Shin, J.: InstaGAN: instance-aware image-to-image translation. ArXiv abs/1812.10889 (2018). https://api.semanticscholar.org/CorpusID:57189269
12. Park, D.Y., Lee, K.H.: Arbitrary style transfer with style-attentional networks. In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 5873–5881 (2018). https://api.semanticscholar.org/CorpusID:54447797
13. Phillips, F.Y., Mackintosh, B.: Wiki art gallery, Inc.: A case for critical thinking. Issues Account. Educ. **26**, 593–608 (2011). https://api.semanticscholar.org/CorpusID:154992634
14. Qin, X., Zhang, Z., Huang, C., Dehghan, M., Zaiane, O.R., Jagersand, M.: U2-Net: going deeper with nested u-structure for salient object detection. Pattern Recogn. **106**, 107404 (2020)
15. Sbai, O., Elhoseiny, M., Bordes, A., LeCun, Y., Couprie, C.: Design: design inspiration from generative networks. In: ECCV Workshops (2018). https://api.semanticscholar.org/CorpusID:4565882
16. Sun, K., Zhang, J., Zhang, P., Yuan, K., Li, G.: TsrNet: a two-stage unsupervised approach for clothing region-specific textures style transfer. J. Vis. Commun. Image Represent. **91**, 103778 (2023). https://api.semanticscholar.org/CorpusID:256714697
17. Ulyanov, D., Lebedev, V., Vedaldi, A., Lempitsky, V.S.: Texture networks: feed-forward synthesis of textures and stylized images. ArXiv abs/1603.03417 (2016). https://api.semanticscholar.org/CorpusID:16728483
18. Vaswani, A., et al.: Attention is all you need. In: Neural Information Processing Systems (2017). https://api.semanticscholar.org/CorpusID:13756489
19. Wang, X., Girshick, R.B., Gupta, A.K., He, K.: Non-local neural networks. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 7794–7803 (2017). https://api.semanticscholar.org/CorpusID:4852647
20. Wen, L., Gao, C., Zou, C.: CAP-VSTNeT: content affinity preserved versatile style transfer. In: 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 18300–18309 (2023). https://api.semanticscholar.org/CorpusID:257900716
21. Yu, C., Hu, Y., Chen, Y., Zeng, B.: Personalized fashion design. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 9046–9055 (2019)
22. Zhang, Y., Li, M., Li, R., Jia, K., Zhang, L.: Exact feature distribution matching for arbitrary style transfer and domain generalization. In: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 8025–8035 (2022). https://api.semanticscholar.org/CorpusID:247450594

Approximate Cuboidization of an Orthogonal Polyhedron: A Combinatorial Approach

Anukul Maity[1], Mousumi Dutt[2](✉), Arindam Biswas[3], and Bhargab B. Bhattacharya[4]

[1] Narula Institute of Technology, Kolkata, India
[2] St. Thomas' College of Engineering and Technology, Kolkata, India
duttmousumi@gmail.com
[3] Indian Institute of Engineering Science and Technology, Shibpur, Howrah, India
[4] Distinguished Visiting Professor (CSE), Indian Institute of Technology, Kharagpur, India

Abstract. A combinatorial algorithm is presented here which partitions a given orthogonal polyhedron, P, (genus zero and non-self-intersecting) into approximately minimum number of cuboids in $O(n \log n)$ time where n is the number of vertices of P. The proposed cuboidization algorithm may start from any rectangular face. The combinatorial rules are formed to determine a cuboid from a given rectangular face. The generated cuboid is removed from the polyhedron, the new faces are created thereof are inserted in a queue. The next rectangular faces of traversal is selected from the queue. If cuboid extraction disconnects the input polyhedron, all the disconnected parts will be tracked through the queue as it stores newly generated faces. The cuboidization algorithm has various applications in 3D modelling and 3D shape analysis.

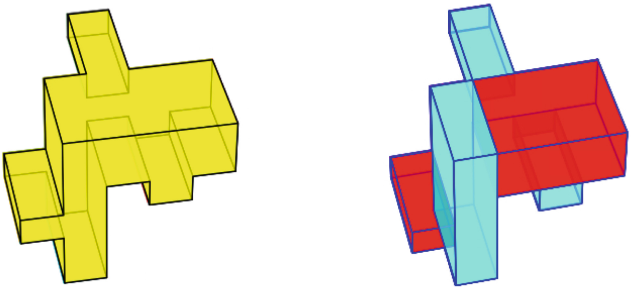
Keywords: Cuboidization · Orthogonal Polyhedron · Polyhedron Decomposition · Minimal Partitioning · Shape Analysis

1 Introduction

Decomposition of complex polyhedron or polygons into meaningful smaller parts is an important area of research in the field of digital geometry. The decomposition can be categorised into two types of problems- covering problem and partitioning problem, where the input polyhedron may be hole-free or with holes [3,9,20]. In 2D, some polynomial time polygon splitting algorithms are studied in [4,8,11,14,17,23,24,27,29,33]. The approximate decomposition of orthogonal polygon is proposed in [12,13]. In [30], minimum decomposition of any arbitrary polygons with holes is proposed which is NP-hard. The 2D partitioning and covering both problems are NP-hard in the orthogonal domain [7]. In general, the task becomes NP-hard in three dimensions, as proved by Dielissen and Kaldewaij [5,10]. In computational geometry, partitioning a geometric

A. Antonacopoulos et al. (Eds.): ICPR 2024, LNCS 15332, pp. 253–269, 2025.
https://doi.org/10.1007/978-3-031-78125-4_18

object or shape into simpler components is a classic problem which has various applications [15,17,19,21,25]. The decomposition of an orthogonal polygon into rectangles (which are related to the straight skeleton) is presented in [26].



(a) Input

(b) Output

Fig. 1. (a) Orthogonal Polyhedron, P. (b) Approximate minimum cuboidization of P, where polyhedron edges are marked in blue color and cuboids are marked with two different colors (red and sky-blue). (Color figure online)

The minimal convex decomposition of polyhedra was proposed by Chazelle in 1981 [6]. The algorithm runs in $O(nN^3)$ time, where n is number of vertices and N is number of reflex edges. It has been found that the decomposition problem is NP-hard when it partitions an orthogonal polyhedron into a minimal number of rectangular boxes [5] or it partitions a 3D-histogram into a minimum number of boxes [18]. In 1992, C. L. Bajaj et al. [2] devised an algorithm to decompose a non-convex polyhedron with arbitrary genus and interior voids in $O(nr^2)$ time, where n is the number of edges and r is the number of reflex edges. In 2018, P. Floderus et al. [18] developed a 4-approximation algorithm to partition 3D histograms in $O(m \log m)$ time, where m is the number of corners in orthogonal polyhedra. The t-approximation algorithm to partition a 3D-histogram into a minimum number of boxes is NP-hard [5] even if the height of histogram is two. In 2007, J. M. Lien and N. M. Amato [22] introduced an additional technique that partitions the polyhedron in a way that minimizes the concavity of the identified features after determining which feature(s) is the most concave in each iteration. It requires $O(n^3 \log n)$ time where n is number of edges. In 2002, S. Svensson et al. [32] presented a distance transform-based technique for nearly convex decomposition. A 3D object is approximated by a set of spheres which are obtained by surface points as stated in [28]. A 3D object is decomposed into smaller parts based on splitting the branches of curve skeleton with less curvature variations [31]. In [34], a 3D mesh is decomposed into almost convex components to minimize the number of components by measuring concavity within a specific threshold value. The simple orthogonal polyhedron can be characterized by graph. In [16], algorithms for constructing orthogonal polyhedra from their graphs is presented.

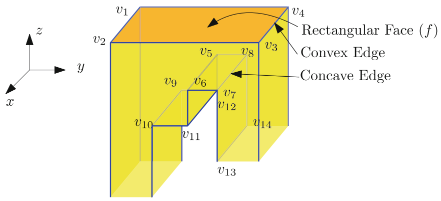


Fig. 2. A part of orthogonal polyhedron and one of its rectangular faces, concave edges, and convex edges are marked.

Here, a combinatorial algorithm is proposed to partition an orthogonal polyhedron (non self-intersecting, genus zero) into approximately minimum number of cuboids in $O(n \log n)$ time, where n is the total number of vertices in the polyhedron. In Fig. 1(a), an orthogonal polyhedron is given and its approximate minimum cuboidization is depicted in Fig. 1(b). The number of vertices of the polyhedron is 36, the number of edges is 50, the number of faces is 18, the number of concave edges is 8, and the number of components is 5 which is minimum. The paper is organized as follows. The definitions and preliminaries are explained in Sect. 2. In Sect. 3, the combinatorial rules, algorithm, time complexity, proof of correctness, and the demonstration are proposed. The experimental results are given in Sect. 4. The concluding remarks are presented in Sect. 5.

2 Definitions and Preliminaries

Definition 1 (Orthogonal Polyhedron). *An orthogonal polyhedron is one all of whose faces meet at right angles, and all of whose edges are parallel to the axes of a Cartesian coordinate system.*

Definition 2 (Simple Orthogonal Polyhedron). *A simple orthogonal polyhedron has axis-parallel faces that are simple polygons and three perpendicular edges incident at each vertex.*

A simple orthogonal polyhedron is genus zero and non-self-intersecting.

Definition 3 (Convex Edge). *If the angle between two faces of an orthogonal polyhedron, P, incident at an edge, e, is* 90° *considering the interior of P, then the edge e is termed as convex edge.*

Definition 4 (Concave Edge). *If the angle between two faces of an orthogonal polyhedron, P, incident at an edge, e, is* 270° *considering the interior of P, then the edge e is termed as concave edge.*

Definition 5 (Convex Vertex and Concave Vertex). *The vertices associated with convex edges are termed as convex vertices and the vertices associated with concave edges are termed as concave vertices.*

In [1], the relationship between convex and concave vertices in an orthogonal polyhedron is described. There are eight more convex vertices compared to concave vertices in an orthogonal polyhedron.

Definition 6 (Cuboid). *A cuboid is a simple orthogonal polyhedron with six faces and twelve convex edges.*

Definition 7 (Rectangular Face). *A rectangular face of an orthogonal polyhedron consists of four edges.*

In Fig. 2, a part of orthogonal polyhedron is shown. A rectangular face, concave edge, and convex edge are shown. All the faces, edges, and vertices of the orthogonal polyhedron are stored in a doubly connected edge list (DCEL). Three lexicographically sorted lists are generated from DCEL, say L_{xy}, L_{yz}, and L_{zx}. Since orthogonal polyhedron is considered, each face has same x or y or z value. For L_{xy}, the face with same z-value are sorted in ascending order of z-value at the first level. At the second level, the faces are sorted w.r.t. x as primary key and y as secondary key. Similarly, L_{yz} and L_{zx} are generated.

3 Procedure to Partition Into Cuboids

The starting rectangular face is selected from L_{xy}. The cuboid is formed based on the combinatorial rules and included in the set of cuboids. The detected cuboid is deducted from the orthogonal polyhedron and the new faces thus formed are inserted into a queue. The next face of traversal is determined from the queue and the procedure continues until the orthogonal polyhedron is reduced to a single cuboid. If the next face of traversal is not a rectangular face, then a rectangular face is selected from that part of the polyhedron using any of the lists L_{xy}, L_{yz}, and L_{zx}. In Sect. 3.1, the rules to obtain a cuboid are presented. The algorithm of partitioning the 3D orthogonal polyhedron into approximate minimum number of cuboids is discussed in Sect. 3.2. The time complexity analysis is presented in Sect. 3.3. Section 3.4 presents a demonstration of cuboidization of an orthogonal polyhedron. In Sect. 3.5, the proof of correctness of the algorithm is proposed.

3.1 Rules for Cuboidization

The rules to detect cuboid is discussed as follows. A rectangular face is selected from the queue (marked in purple in Fig. 3 and Fig. 4). The face may lie in any one of the plane, xy-plane, yz-plane, or zx-plane. If the rectangular face, f, is in xy-plane, L_{xy} is traversed to find the faces which are obtained while sweeping f through the interior of the polyhedron. If f is in yz-plane or zx-plane, L_{yz} and L_{zx} are traversed respectively. The rectangular face f is swept until it gets an obstacle which is one or more faces. In Fig. 2, the obstacle for the rectangular face, f is the face formed by the vertices $\{v_5, v_6, v_7, v_8\}$.

Rule-1: This rule is applied when f (shown in purple color) is swept until it hits an obstacle (face(s) f_1 and/or f_2) and correspondingly the cuboid is

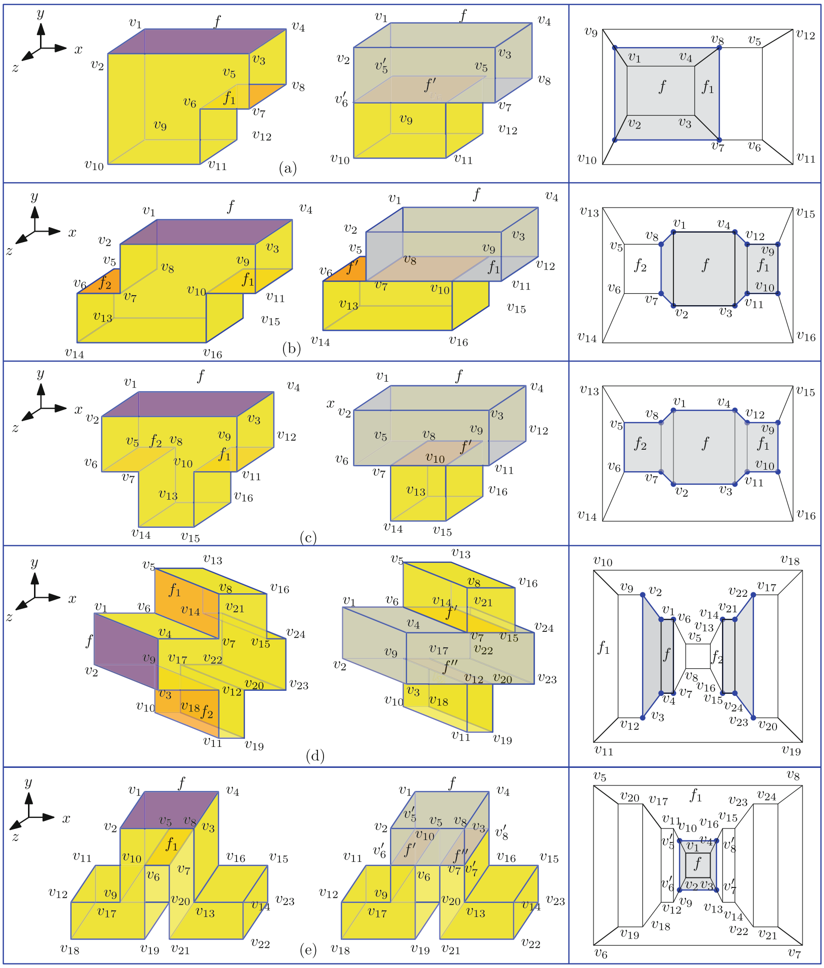

y
x
z
(a)
v_1 f v_4 v_2 v_3 v_5 v_8 v_6 f_1 v_7 v_9 v_{12} v_{10} v_{11}
v_1 f v_4 v_2 v_5' v_3 v_5 v_8 f' v_6' v_7 v_9 v_{12} v_{10} v_{11}
v_9 v_8 v_5 v_{12} v_1 v_4 f f_1 v_2 v_3 v_7 v_6 v_{10} v_{11}
(b)
v_1 f v_4 v_2 v_5 v_8 v_9 v_3 v_6 f_2 v_7 v_{10} f_1 v_{11} v_{13} v_{15} v_{14} v_{16}
v_1 f v_4 v_2 v_3 v_5 v_9 v_6 f' v_8 f_1 v_{12} v_7 v_{10} v_{11} v_{13} v_{15} v_{14} v_{16}
v_{13} v_{15} v_8 v_1 v_4 v_{12} v_5 v_9 f_2 f f_1 v_6 v_{10} v_7 v_{11} v_2 v_3 v_{14} v_{16}
(c)
v_1 f v_4 v_2 v_3 v_{12} v_5 f_2 v_8 v_9 v_6 v_{10} f_1 v_7 v_{11} v_{13} v_{16} v_{14} v_{15}
v_1 f v_4 x v_2 v_3 v_{12} v_5 v_8 v_9 f' v_6 v_{10} v_7 v_{11} v_{13} v_{16} v_{14} v_{15}
v_{13} v_{15} v_1 v_4 v_8 v_{12} v_5 v_9 f_2 f f_1 v_6 v_{10} v_7 v_{11} v_2 v_3 v_{14} v_{16}
(d)
v_5 v_{13} f_1 v_8 v_{16} v_1 v_6 v_{14} v_{21} f v_4 v_7 v_{15} v_{24} v_9 v_{17} v_{22} v_2 v_{20} v_{23} v_3 v_{12} v_{10} v_{18} f_2 v_{11} v_{19}
v_5 v_{13} v_8 v_{16} v_1 v_6 v_{14} f' v_{21} v_4 v_{24} v_7 v_{15} v_9 v_{17} v_{22} v_2 f'' v_{12} v_{20} v_{23} v_3 v_{18} v_{10} v_{11} v_{19}
v_{10} v_{18} v_9 v_2 v_{22} v_{17} v_1 v_6 v_{14} v_{21} v_{13} v_5 f_1 f f_2 v_8 v_{16} v_7 v_{15} v_4 v_{24} v_3 v_{23} v_{12} v_{20} v_{11} v_{19}
(e)
v_1 f v_4 v_2 v_5 v_8 v_3 f_1 v_{11} v_{10} v_{16} v_{15} v_6 v_7 v_{12} v_9 v_{20} v_{14} v_{23} v_{17} v_{13} v_{18} v_{19} v_{21} v_{22}
v_1 f v_4 v_5' v_2 v_5 v_8 v_3 v_8' v_{10} v_6' f' f'' v_{11} v_6 v_7 v_7' v_{16} v_{15} v_{12} v_9 v_{20} v_{14} v_{23} v_{17} v_{13} v_{18} v_{19} v_{21} v_{22}
v_5 v_8 v_{20} v_{17} f_1 v_{23} v_{24} v_{11} v_{10} v_{16} v_{15} v_5' v_1 v_4 v_8' f v_6' v_2 v_3 v_7' v_{12} v_9 v_{13} v_{14} v_{18} v_{22} v_{19} v_{21} v_6 v_7

Fig. 3. Rule-1 is illustrated. The polyhedron (P) is marked in yellow color, convex face (f) is marked in purple color and the removed cuboid is marked in gray color. f' (orange color) is inserted into the queue. The planar graph of the polyhedron is shown at the right column where the extracted cuboid is in gray color. (Color figure online)

determined from f to obstacle face(s) (shown in gray color) (see Fig. 3). The different cases are shown in Fig. 3. The extraction of cuboid creates one or more new faces (f' and f'' in Fig. 3) and they are included in the queue.

RULE-2: This rule is applied when f (shown in purple color) is swept up to a rectangular face, f_1, where all four or three edges are concave edges. The corresponding cuboid is determined from f to f_1 (shown in gray color) (see Fig. 4). The extraction of cuboid creates one new face f' and it is included in the queue.

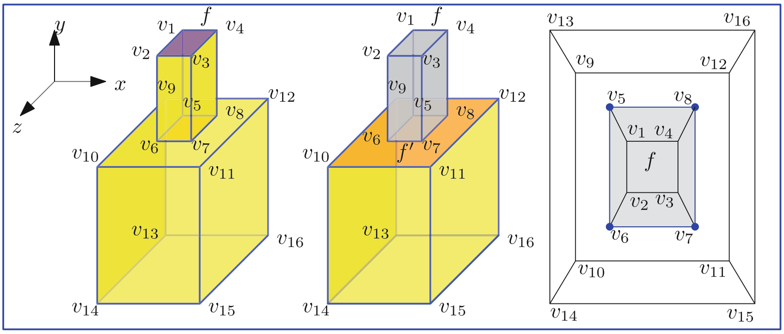


Fig. 4. Rule-2 is illustrated. The polyhedron (P) is marked in yellow color, convex face (f) is marked in purple color and the cuboid is marked in gray color. f' (orange color) is inserted into the queue. The planar graph of the polyhedron is shown at the right column where the extracted cuboid is in gray color. (Color figure online)

In case of Rule 2, if the face f is swept up to the obstacle to extract the cuboid, the remaining polyhedron will not remain genus-zero. It is to be noted here that the inserted faces in the queue may not be always rectangular. The polyhedron can be represented by planar graph. The extracted cuboids are marked in gray color in the planar graph.

3.2 Algorithm for Cuboidization

The algorithm 3D-Cuboidization finds approximate minimum partition of orthogonal polyhedron P. The DCEL, L, of P is taken as input. The output is the list of cuboids, L_c. Step 1 finds lexicographically sorted lists, L_{xy}, L_{yz}, L_{zx} w.r.t. each plane by calling the procedure Sort-List which takes the DCEL, L, as input. In step 2, L_{xy} is scanned to find a rectangular face which is the start face (say, f) of determining cuboidization (procedure Find-Start-Face is called). The queue, Q and the list of cuboids, L_c, are initially empty (step 3). The start face, f, is inserted in the queue, Q (step 4). Steps 5–21, the cuboids are determined. The loop runs until the Q is empty. If the face in the front of Q is not rectangular (checked by the procedure Check-Rect), then a rectangular face is found from that part of the polyhedron using the lists L_{xy}, L_{yz}, or L_{zx} by the procedure Find-Rect-Face and inserted in Q (steps 6–7). The rectangular face f is extracted from Q (step 8). The face, f, may lie in anyone of the plane, which is detected by the procedure Face-Alignment and the value is assigned to t (step 9). If f is in xy-plane, $t = 0$. $t = 1$ when f is in yz-plane. Otherwise, $t = 2$, i.e., when f is in zx-plane. Based on the values of t, obstacle faces are found using the procedure Find-Obstacle and those faces are stored in L_{obs} (step 10–15). In Sect. 3.1, the obstacle faces are discussed with figures. If four or three edges of obstacle face are concave edges (detected by the procedure Check-Edge), Rule 2 is used (steps 16–17), otherwise Rule 1 is applied (steps 18–19). The procedure Apply-Rule2 is used and the extracted cuboid

Algorithm 1: 3D-Cuboidization

```
   Input: DCEL of the given polyhedron, $L$
   Output: List of Components, $L_c$
1  $L_{xy}, L_{yz}, L_{zx} \leftarrow$ Sort-List($L$)
2  $f \leftarrow$ Find-Start-Face($L_{xy}$)
3  $Q \leftarrow \emptyset, L_c \leftarrow \emptyset$
4  $Q \leftarrow Q \cup \{f\}$
5  while $Q \neq \emptyset$ do
6      if Check-Rect(Head($Q$)) =false then
7          Find-Rect-Face($Q, L_{xy}, L_{yz}, L_{zx}$)
8      $f \leftarrow$ DEQUEUE($Q$)
9      $t \leftarrow$ Face-Alignment($f$)
10     if $t = 0$ then
11         $L_{obs} \leftarrow$ Find-Obstacle($f, L_{xy}$)
12     else if $t = 1$ then
13         $L_{obs} \leftarrow$ Find-Obstacle($f, L_{yz}$)
14     else
15         $L_{obs} \leftarrow$ Find-Obstacle($f, L_{zx}$)
16     if Check-Edge($L_{obs}$) = true then
17         $c, L_f \leftarrow$ Apply-Rule2($f, L_{obs}, t, L_{xy}, L_{yz}, L_{zx}$)
18     else
19         $c, L_f \leftarrow$ Apply-Rule1($f, L_{obs}, t, L_{xy}, L_{yz}, L_{zx}$)
20     $Q \leftarrow$ ENQUEUE($Q, L_f$)
21     $L_c \leftarrow L_c \cup \{c\}$
22 return $L_c$
```

is assigned to c (step 17). After extraction of the cuboid, the next face as per Rule 2 is inserted in to the list L_f (step 17). In steps 18–19, Rule 1 is applied by calling the procedure Apply-Rule1. After extraction of the cuboid, the lists L_{xy}, L_{yz}, and L_{zx} are updated. The next faces are inserted into Q (step 20) and the extracted cuboid is included in the list of cuboids, L_c (step 21). The total list of cuboids are returned when the loop terminates (step 21).

3.3 Time Complexity Analysis

Let n and f be the total number of vertices and faces in the orthogonal polyhedron respectively. To create the lexicographically sorted lists of faces L_{xy}, L_{yz}, and L_{zx}, $O(f \log f)$ time is needed. The procedure Find-Start-Face will traverse sequentially the list L_{xy} and takes linear time w.r.t. the number of faces (i.e., $O(f)$). It is to be noted here that the faces of polyhedron which are in xy-plane are in L_{xy}. The loop will traverse linearly w.r.t. the number of rectangular faces (i.e., $O(f)$). To check whether there is a rectangular face at the front of the queue, Q, it takes constant time. When a face is rectangular, there are only four vertices and four convex edges. If the extracted face from Q is not

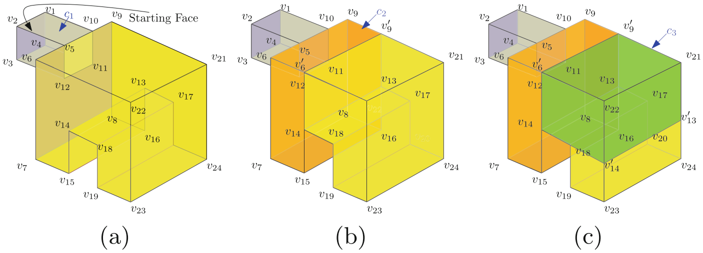


Fig. 5. Demonstration of cuboidization of a simple orthogonal polyhedron.

rectangular, then a rectangular face is found from that part of the polyhedron using the lexicographically sorted lists. Thus, to find next rectangular face takes $O(f)$ time. The procedures ENQUEUE and DEQUEUE take constant time. The alignment of face can be determined in constant time by checking the coordinates of the vertices in the face. The procedure FIND-OBSTACLE traverses a part of any of the lexicographically sorted lists, which is linear w.r.t. the total number of faces in the corresponding plane. To check whether four or three edges at the obstacle face are concave, the procedure CHECK-EDGE checks face and edge lists, which needs linear time (linear w.r.t. the number of edges of the corresponding faces). The procedures APPLY-RULE1 and APPLY-RULE2 take linear time. The lists, L_{xy}, L_{yz}, and L_{zx} are updated in $O(f)$ time. Since, $f < n$ in an orthogonal polyhedron, $O(f) < O(n)$. Thus, total time complexity is $O(n \log n)$ time.

3.4 Demonstration of Cuboidization

In Fig. 5, a demonstration is illustrated. The starting face is indicated by an arrow and the first cuboid, c_1 is extracted based on Rule 2 (Fig. 5(a)). The next face of traversal is determined and next cuboid, c_2, is extracted based on Rule 1 (Fig. 5(b)). Again, the next face of traversal is determined and based on Rule 1 the cuboid c_3 is extracted (Fig. 5(c)). The rest of the polyhedron is a cuboid and the algorithm terminates.

The algorithm is starting face dependent as shown in Fig. 6. For the one orthogonal polyhedron, the algorithm is applied from two different start faces as shown in Fig. 6(a) and Fig. 6(b). In Fig. 6(a), the minimum number of components are not obtained whereas in Fig. 6(b), the minimum number of components are obtained.

3.5 Proof of Correctness

The total number of cuboids depend on the number of concave edges.

Lemma 1. *The maximum number of cuboids which can be extracted from an orthogonal polyhedron (genus zero and non-self-intersecting) is one more than the total number of concave edges in the orthogonal polyhedron.*

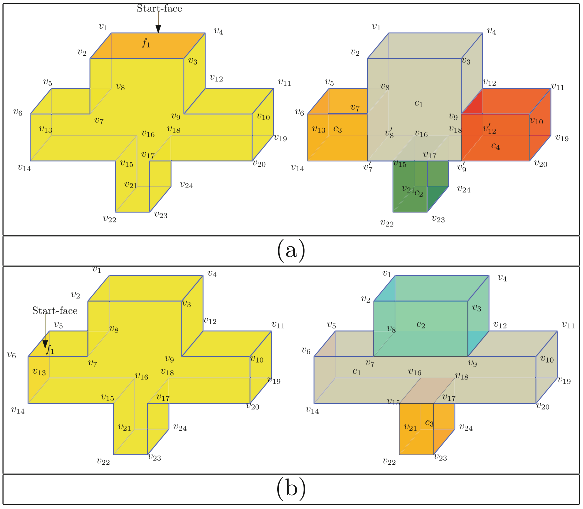

Start-face
f_1
c_1
c_2
c_3
c_4
v_1 v_2 v_3 v_4 v_5 v_6 v_7 v_8 v_9 v_{10} v_{11} v_{12}
v_{13} v_{14} v_{15} v_{16} v_{17} v_{18} v_{19} v_{20} v_{21} v_{22} v_{23} v_{24}
v'_7 v'_8 v'_9 v'_{12}
(a)
(b)

Fig. 6. A demonstration of approximate minimum cuboidization.

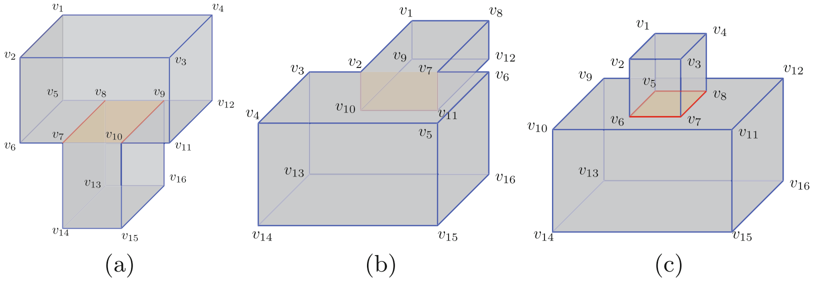

v_1 v_2 v_3 v_4 v_5 v_6 v_7 v_8 v_9 v_{10} v_{11} v_{12} v_{13} v_{14} v_{15} v_{16}
(a)
(b)
(c)

Fig. 7. Illustration of net concave edges. The concave edges are marked by red color. (a) There is one net concave edge out of the two concave edges (v_7v_8) and $(v_9v_{10}(Colorfigureonline))$, (b) There is one net concave edge out of the three concave edges (v_7v_{11}), $(v_{10}v_{11})$, and (v_2v_{10}), (c) There is one net concave edge out of the four concave edges (v_8v_5), (v_5v_6), (v_6v_7), and (v_7v_8).

Proof. Let k be the total number of concave edges in a given orthogonal polyhedron (genus zero and non-self-intersecting). For each of the concave edges, one cuboid is extracted. Whenever a cuboid is extracted no more concave edges are generated in the residual polyhedron. Thus, at last there will not be any concave edge in the residual polyhedron. When a polyhedron does not have any concave edge, the polyhedron will contain only convex edges. A polyhedron with convex edges only must be a cuboid. At last the rest of the polyhedron will be there as the last cuboid. Thus, total number of cuboids is one more than the total number of concave edges. □

If the two concave edges, e_i and e_j, are at the same plane (along any of the three planes) and e_i and e_j can be connected by a rectangular face (which is not a polyhedron face) such that the face totally lies within the polyhedron and the face can move along the normal of the said plane, then one of the edges is counted as net concave edge and another is discarded as joining the two concave edges extract one cuboid instead of two. If there are three or four concave edges forming a rectangular face which can move along the normal of the said plane, then out of three or four concave edges only one is taken as net concave edge and the others are discarded. These three or four concave edges extract one cuboid (see Fig. 4). The net concave edges are shown in Fig. 7. If the rectangular face formed by the concave edges as said above (see Fig. 7) is in xy-plane, then the face can sweep along z-axis. Other concave edges which cannot form a rectangular face as said above, are counted as one as net concave edge.

Lemma 2. *The minimum number of cuboids which can be extracted from an orthogonal polyhedron (genus zero and non-self-intersecting) is one more than the total number of net concave edges.*

Proof. When two or three or four concave edges are in the same plane such that they can be connected by a rectangular face and they can be counted as one net concave edge, then one cuboid can be extracted for those concave edges considered together. Let k' be the net concave edges in the orthogonal polyhedron. If each cuboid in the orthogonal polyhedron is extracted per net concave edge, then rest of the polyhedron contains no more concave edges but only convex edges. It is to be noted here that whenever a cuboid is extracted no more concave edges are generated in the extracted cuboid or in the residual polyhedron. When there are only convex edges, the corresponding polyhedron is a cuboid. It can be said that the minimum number of cuboids extracted is one more than the total number of net concave edges as rest of the polyhedron extracts the last cuboid. □

Let k be the total number of concave edges and k' be the total number of net concave edges. The approximation ratio is $\frac{k'+1}{k+1} \simeq \frac{k'}{k}$ which is less than or equal to one as $k' \leq k$. The algorithm provides approximate minimum number of cuboids for the given polyhedron. As the algorithm is dependent on the starting face, the set of concave edges belonging to a single net concave edge may be considered separately. This increases the count of extracted components but lie within the range $[k'+1, k+1]$.

Theorem 1. *If k' is the net concave edges, then in the worst case maximum and minimum number of components are $4 \times k' + 1$ and $k' + 1$ respectively.*

Proof. If k' is the net concave edges then $k' + 1$ is the minimum number of components as proved in Lemma 2. As per Lemma 2, four or three or two concave edges may be considered as one net concave edge. At the worst case, one net concave edge is equivalent to four concave edges. The total number of concave edges is four times of net concave edges in the worst case. Thus, maximum number of components in an orthogonal polyhedron is one more than the total number concave edges as per Lemma 1. Hence maximum number of components in the orthogonal polyhedron is $4 \times k' + 1$. □

All the cuboids extracted reconstructs the orthogonal polyhedron. The extracted cuboids do not overlap. Only common face(s) or part of face(s) is shared. The whole polyhedron is traversed to extract the cuboids. When a cuboid is extracted, it is added in the list of cuboids, C. If the cuboid c_i is extracted, then $C = C \cup c_i$ and the orthogonal polyhedron P becomes $P \setminus C$. When all the cuboids are extracted $P = C$. The algorithm is designed for orthogonal cases only. For general polyhedrons, the algorithm needs to be modified.

Theorem 2. *The algorithm* 3D-Cuboidization *terminates properly and generates correct results.*

Proof. Whenever a cuboid is extracted one or more new faces are created and those faces are inserted in a queue. Thus, if the polyhedron gets disconnected after the extraction of a cuboid, no parts of the polyhedron are left out for the traversal as per the algorithm. Each face is retrieved from the queue and the corresponding rule is applied to extract the cuboid. Accordingly, new faces are inserted in the queue. If the new face in the queue is non-rectangular, a rectangular face is searched from that part of the polyhedron. When the queue contains only one face and the remaining polyhedron is only a cuboid, the algorithm terminates after considering the last cuboid. It can be stated here that the algorithm traverses the whole polyhedron. The extraction of cuboid does not generate a new concave edge. The combinatorial rules are formed in such a way that cuboids are extracted from concave edge(s). Thus the total number of generated cuboids lies between the maximum and minimum range. Rule 2 extracts cuboids up to the face constructed by net concave edges. Thus, after extraction of the cuboid the remaining polyhedron remains genus-zero and non self-intersecting. Hence when the algorithm terminates, it generates correct result. □

4 Experimental Results

The experimental results are generated using Python programming language (Python 3.6.5) on a computer system with Intel core i5 processor and OS Ubuntu 16.04. Some of the experimental results of the cuboidization on different orthogonal polyhedra are given in Fig. 9, Fig. 10 and Fig. 11. Cuboidization is useful for shape analysis. Several parameters of cuboidization can be useful shape descriptors of 3D objects. The data related to experimental results are shown in Table 1 which are the number of vertices, edges, faces, concave edges, convex edges, components as per the given algorithm, and net concave edges. The maximum and minimum number of components can be derived from the above mentioned data as discussed in Sect. 3.5. It has been observed that the generated results in Fig. 9, Fig. 10 and Fig. 11 give the minimum number of cuboids. If there is more difference in the number of concave edges and net concave edges, then the shape of the polyhedron is complex. Depending on the start face, the number of extracted components may vary but lie within the range between the maximum and minimum number of components for the polyhedron. In Fig. 8 a demonstration of approximate minimum cuboidization is shown. There are five components in total which can be glued again to make the original polyhedron. The approximate minimum cuboidization can be applied in different fields like additive manufacturing, gluing, 3D printing, etc.

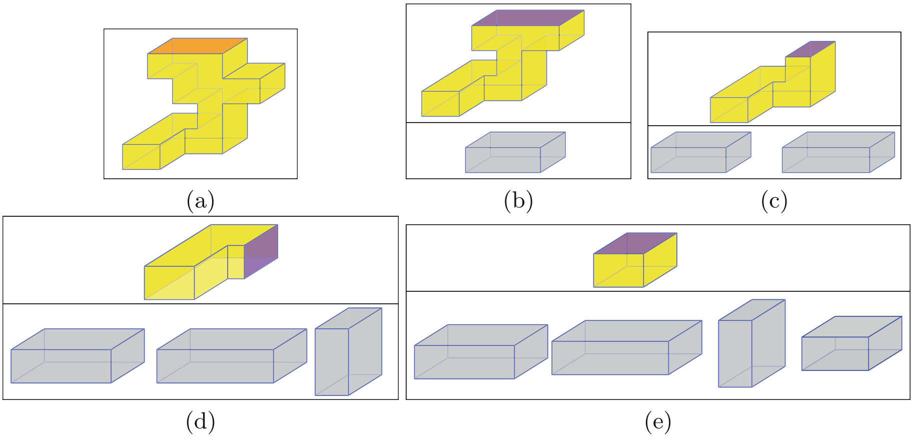


Fig. 8. A demonstration of approximate minimum cuboidization.

Table 1. The data of experimental results shown in Fig. 10

Polyhedron	**Vertices**	**Edges**	**Faces**	**Convex edges**	**Concave edges**	**Net concave edges**	**Components**
Fig. 9(a)	24	36	14	31	5	2	3
Fig. 9(b)	28	42	16	36	6	4	5
Fig. 9(c)	32	46	18	39	7	4	5
Fig. 9(d)	44	58	24	48	10	5	6
Fig. 9(e)	24	36	14	31	5	3	4
Fig. 9(f)	32	46	16	38	8	4	5
Fig. 9(g)	48	70	22	60	10	6	7
Fig. 9(h)	24	32	14	28	4	3	4
Fig. 9(i)	24	36	14	32	4	3	4
Fig. 10(a)	36	54	20	45	9	4	5
Fig. 10(b)	26	39	15	34	5	3	4
Fig. 10(c)	30	40	17	34	6	4	5
Fig. 10(d)	38	50	18	42	8	5	6
Fig. 10(e)	44	60	22	15	9	7	8
Fig. 10(f)	38	50	17	40	10	5	6
Fig. 10(g)	30	40	17	33	7	4	5
Fig. 10(h)	34	44	18	36	8	5	6
Fig. 10(i)	38	56	18	47	9	6	7
Fig. 11(a)	94	128	54	106	22	16	17
Fig. 11(b)	104	156	54	130	26	18	19
Fig. 11(c)	280	356	138	292	64	30	31
Fig. 11(d)	288	384	142	315	69	37	38

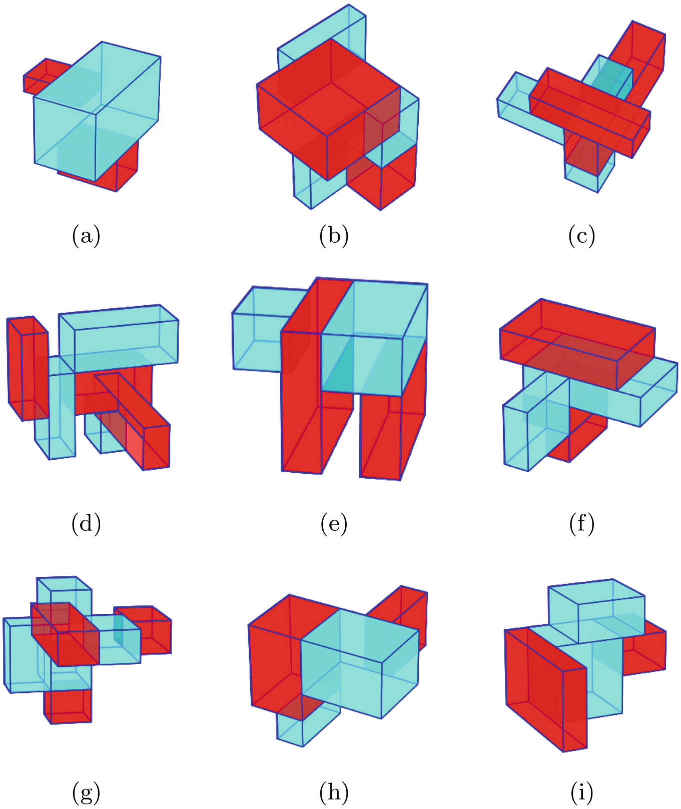
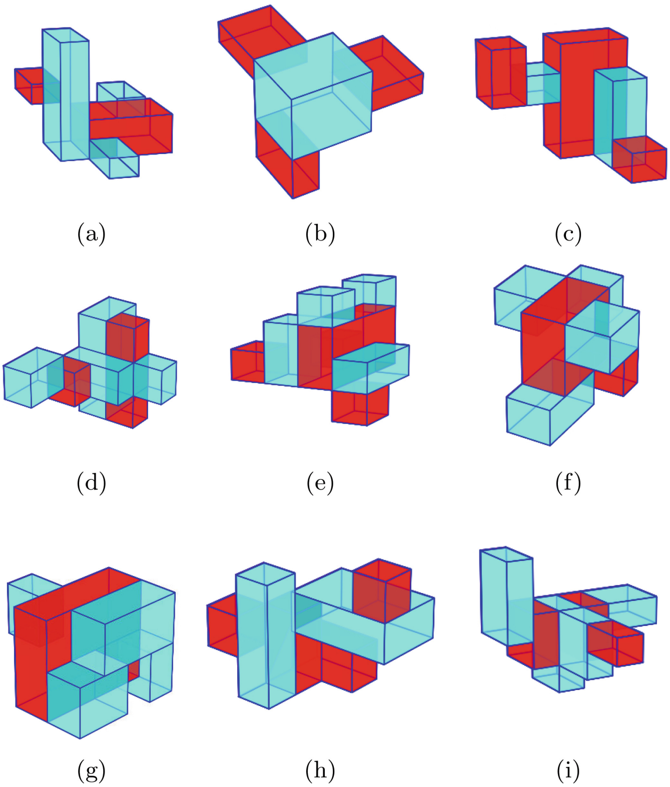


Fig. 9. Experimental results of Cuboidization for a set of orthogonal polyhedrons.


(a) (b) (c)

(d) (e) (f)

(g) (h) (i)

Fig. 10. Experimental results of Cuboidization for another set of orthogonal polyhedrons.

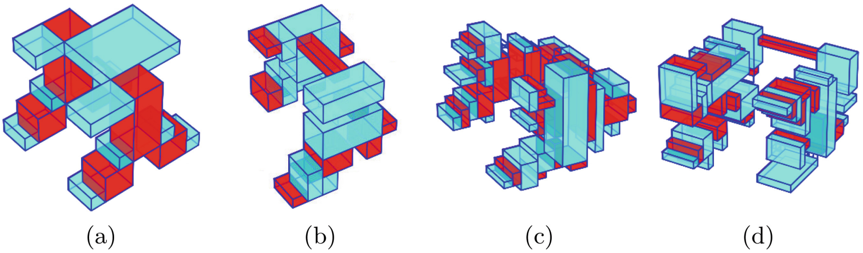

(a) (b) (c) (d)

Fig. 11. Experimental results of Cuboidization for another set of orthogonal polyhedrons.

5 Conclusion

This paper presents a combinatorial algorithm to decompose an orthogonal polyhedron (genus zero and non-self-intersecting) into approximately minimum number of cuboids in $O(n \log n)$ time, where n is the number of vertices of the polyhedron. The combinatorial rules are formed to extract cuboids at each step. The algorithm is not starting point invariant. The approximation ratio is stated in the paper. The demonstration and proof of correctness are presented here. The experimental results show the efficacy of the algorithm. In future, the algorithm can be used for 3D shape analysis, shape retrieval, shape matching, etc. The approximate minimum cuboidization has applications in various fields.

References

1. Aldana-Galván, I., Álvarez-Rebollar, J., Catana-Salazar, J., Jiménez-Salinas, M., Solís-Villarreal, E., Urrutia, J.: Minimizing the solid angle sum of orthogonal polyhedra. Inf. Process. Lett. **143**, 47–50 (2019)
2. Bajaj, C.L., Dey, T.K.: Convex decomposition of polyhedra and robustness. SIAM J. Comput. **21**(2), 339–364 (1992)
3. Berg, M.D., Cheong, O., Kreveld, M.V., Overmars, M.: Computational Geometry: Algorithms and Applications. Springer-Verlag, Heidelberg, 3rd edn. (2008). https://doi.org/10.1007/978-3-540-77974-2
4. Berg, M.D., Kreveld, M.V.: Rectilinear decompositions with low stabbing number. Inf. Process. Lett. **52**(4), 215–221 (1994)
5. Biedl, T., Derka, M., Irvine, V., Lubiw, A., Mondal, D., Turcotte, A.: Partitioning orthogonal histograms into rectangular boxes. In: Bender, M., Farach-Colton, M., Mosteiro, M. (eds.) Proceedings of the LATIN 2018: Theoretical Informatics, vol. 10807, pp. 146–160. Springer International Publishing, Cham (2018). https://doi.org/10.1007/978-3-319-77404-6_12
6. Chazelle, B.: Convex decompositions of polyhedra. In: Proceedings of the 13th Annual ACM Symposium on Theory of Computing, pp. 70–79. STOC'81, Association for Computing Machinery (1981)
7. Culberson, J.C., Reckhow, R.A.: Covering polygons is hard. In: Proceedings of 29th Annual Symposium on Foundations of Computer Science, pp. 601–611. IEEE Computer Society, White Plains, NY, USA (1988)
8. Demir, İ, Aliaga, D.G., Benes, B.: Near-convex decomposition and layering for efficient 3D printing. Addit. Manuf. **21**, 383–394 (2018)
9. Devadoss, S.L., Rourke, J.O.: Discrete and Computational Geometry. Princeton University Press (2011)
10. Dielissen, V.J., Kaldewaij, A.: Rectangular partition is polynomial in two dimensions but NP-complete in three. Inf. Process. Lett. **38**(1), 1–6 (1991)
11. Durocher, S., Mehrabi, S.: Computing conforming partitions of orthogonal polygons with minimum stabbing number. Theoret. Comput. Sci. **689**, 157–168 (2017)
12. Dutt, M., Biswas, A., Bhowmick, P.: ACCORD: with approximate covering of convex orthogonal decomposition. In: Debled-Rennesson, I., Domenjoud, E., Kerautret, B., Even, P. (eds.) Proceeding of the 16th IAPR International Conference on Discrete Geometry for Computer Imagery, (DGCI). Lecture Notes in Computer Science, vol. 6607, pp. 489–500. Springer, Nancy, France (2011)

13. Dutt, M., Biswas, A., Bhowmick, P.: Approximate partitioning of 2D objects into orthogonally convex components. Comput. Vis. Image Underst. **117**(4), 326–341 (2013)
14. Eppstein, D.: Graph-theoretic solutions to computational geometry problems. In: Paul, C., Habib, M. (eds.) Proccedings of the International Workshop on Graph-Theoretic Concepts in Computer Science. WG 2009, vol. 5911, pp. 1–16. Springer, Berlin Heidelberg, Berlin, Heidelberg (2009). https://doi.org/10.1007/978-3-642-11409-0_1
15. Eppstein, D.: Orthogonal dissection into few rectangles. Discrete Comput. Geom., 1–20 (2023). https://doi.org/10.1007/s00454-023-00614-w
16. Eppstein, D., Mumford, E.: Steinitz theorems for simple orthogonal polyhedra. J. Comput. Geom. **5**(1), 179–244 (2014)
17. Ferrari, L., Sankar, P.V., Sklansky, J.: Minimal rectangular partitions of digitized blobs. Comput. Vis. Graph. Image Process. **28**(1), 58–71 (1984)
18. Floderus, P., Jansson, J., Levcopoulos, C., Lingas, A., Sledneu, D.: 3D rectangulations and geometric matrix multiplication. Algorithmica **80**, 136–154 (2018)
19. Hannenhalli, S., Hubell, E., Lipshutz, R., Pevzner, P.: Combinatorial algorithms for design of DNA arrays. Adv. Biochem. Eng. Biotechnol. **77**, 1–19 (2002)
20. Klette, R., Rosenfeld, A.: Digital Geometry: Geometric Methods for Digital Picture Analysis. Morgan Kaufmann, San Francisco (2004)
21. Li, G., Zhang, H.: A rectangular partition algorithm for planar self-assembly. In: International Conference on Intelligent Robots and Systems, pp. 3213 – 3218. 2005 IEEE/RSJ, IEEE (2005)
22. Lien, J.M., Amato, N.M.: Approximate convex decomposition of polyhedra and its applications. Comput. Aided Geometric Des. **25**(7), 503–522 (2008)
23. Lingas, A., Soltan, V.: Minimum convex partition of a polygon with holes by cuts in given directions. In: Asano, T., Igarashi, Y., Nagamochi, H., Miyano, S., Suri, S. (eds.) Proceedings of 7th International Symposium on Algorithms and Computation (ISAAC 2006). vol. 1178, pp. 315–325. (LNCS), Springer, Heidelberg (1996). https://doi.org/10.1007/BFb0009508
24. Lingas, A., Soltan, V.: Minimum convex partition of a polygon with holes by cuts in given directions. Theory Comput. Syst. **31**, 507–538 (1998)
25. Lopez, M.A., Mehta, D.P.: Efficient decomposition of polygons into L-shapes with application to VLSI layouts. ACM Trans. Des. Autom. Electron. Syst. **1**(3), 371–395 (1996)
26. Maity, A., Dutt, M., Biswas, A.: Rectangularization of digital objects and its relation with straight skeletons. In: Barneva, R.P., Brimkov, V.E., Nordo, G. (eds.) Proceedings of Combinatorial Image Analysis, vol. 13348, pp. 31–45. Springer International Publishing, Cham (2023). https://doi.org/10.1007/978-3-031-23612-9_2
27. Nahar, S., Sahni, S.: Fast algorithm for polygon decomposition. IEEE Trans. Comput. Aided Des. Integr. Circuits Syst. **7**(4), 473–483 (1988)
28. O'Rourke, J., Badler, N.: Decomposition of three-dimensional objects into spheres. IEEE Trans. Pattern Anal. Mach. Intell. **1**(3), 295–305 (1979)
29. O'Rourke, J., Tewari, G.: The structure of optimal partitions of orthogonal polygons into fat rectangles. Comput. Geom. **28**(1), 49–71 (2004)
30. Rourke, J.O., Supowit, K.J.: Some NP-hard polygon decomposition problems. IEEE Trans. Inf. Theory **29**(2), 181–190 (1983)
31. Serino, L., Arcelli, C., di Baja, G.S.: Decomposing 3D objects in simple parts characterized by rectilinear spines. Int. J. Pattern Recognit Artif Intell. **28**(07), 1460010 (2014)

32. Svensson, S., Di Baja, G.S.: Using distance transforms to decompose 3D discrete objects. Image Vis. Comput. **20**(8), 529–540 (2002)
33. Thul, D., Ladický, L., Jeong, S., Pollefeys, M.: Approximate convex decomposition and transfer for animated meshes. ACM Transactions on Graphics **37**(6), 226 (2018)
34. Wei, X., Liu, M., Ling, Z., Su, H.: Approximate convex decomposition for 3D meshes with collision-aware concavity and tree search. ACM Trans. Graph. **41**(4), 1–18 (2022)

Enhancing Multi-exposure High Dynamic Range Imaging with Overlapped Codebook for Improved Representation Learning

Keuntek Lee[1], Jaehyun Park[2], and Nam Ik Cho[1,2](✉)

[1] Department of ECE, INMC, Seoul National University, Seoul, Korea
{leekt000,nicho}@snu.ac.kr

[2] IPAI, INMC, Seoul National University, Seoul, Korea
jaep970805@gmail.com

Abstract. High dynamic range (HDR) imaging technique aims to create realistic HDR images from low dynamic range (LDR) inputs. Specifically, Multi-exposure HDR imaging uses multiple LDR frames taken from the same scene to improve reconstruction performance. However, there are often discrepancies in motion among the frames, and different exposure settings for each capture can lead to saturated regions. In this work, we first propose an Overlapped codebook (OLC) scheme, which can improve the capability of the VQGAN framework for learning implicit HDR representations by modeling the common exposure bracket process in the shared codebook structure. Further, we develop a new HDR network that utilizes HDR representations obtained from a pre-trained VQ network and OLC. This allows us to compensate for saturated regions and enhance overall visual quality. We have tested our approach extensively on various datasets and have demonstrated that it outperforms previous methods both qualitatively and quantitatively.

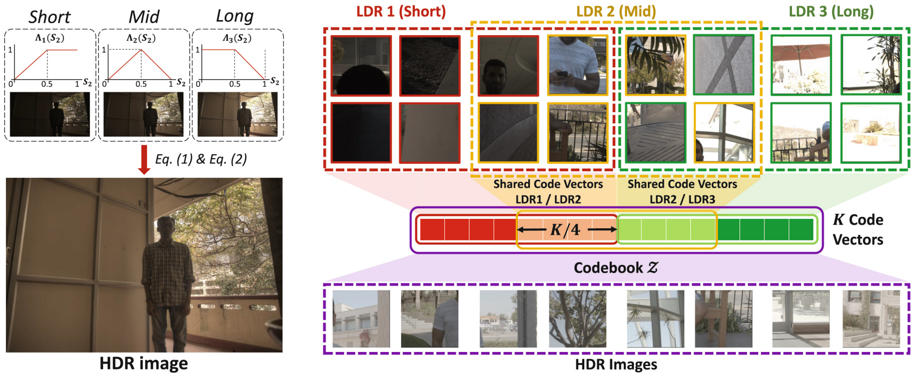
Keywords: Exposure fusion · HDR imaging · Vector quantization

1 Introduction

The task of multi-exposure high dynamic range (HDR) imaging is to create a high-quality HDR image from multiple low dynamic range (LDR) images that were taken with different exposure settings. This approach is superior to single-image HDR imaging, which lacks information and produces lower-quality results. By utilizing more information from multiple frames when LDR frames are perfectly still, multi-exposure HDR imaging can produce finer HDR results. However, LDR frames taken by exposure bracketing have motion differences from

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/978-3-031-78125-4_19.

A. Antonacopoulos et al. (Eds.): ICPR 2024, LNCS 15332, pp. 270–286, 2025.
https://doi.org/10.1007/978-3-031-78125-4_19



(a) Exposure bracketing (b) Proposed Overlapped Codebook (OLC) scheme

Fig. 1. Illustration of (a) conventional exposure bracketing process with triangle function $(\Lambda_1, \Lambda_2, \Lambda_3)$ and (b) proposed Overlapped Codebook (OLC) scheme for multi-exposure HDR imaging. The proposed OLC scheme is able to represent HDR images with a combination of LDR representations by aligning the exposure bracket process with its codebook structure.

each other, and each LDR image has over- or under-exposed regions, which can lead to undesirable artifacts such as ghosting and washed-out areas in the final HDR image. To deal with these issues, earlier works [7,8,22] used pre-processing steps to align the LDR frames before merging them, by using optical flow or homography transformation. However, such explicit alignment methods can have estimation errors, bringing misaligned frames to the following merging stage.

Recently, convolutional neural networks (CNNs) have achieved notable successes in various computer vision areas, including HDR imaging. Kalantari *et al.* [7] first proposed a CNN-based merging network for multi-exposure HDR imaging. Yan *et al.* [9] proposed an attention-based network that implicitly aligns non-reference frames at the feature level. More recently, Niu *et al.* [11] proposed an HDR method based on the generative adversarial network (GAN) [10], and Liu *et al.* [16] presented an algorithm based on the Vision Transformer (ViT) [12]. Although CNN-based methods generally outperform traditional methods in HDR reconstruction, they still struggle with saturated regions and missing details on severely under-/over-exposed LDR frames.

In this work, we introduce a novel HDR reconstruction network with a dual-decoder structure that leverages learned HDR representations to restore fine details and compensate for saturated regions. Our approach employs a vector quantization (VQ) mechanism for learning HDR image representations, specifically proposing the Overlapped Codebook (OLC) scheme that models the exposure bracket fusing process (Fig. 1(a)). The proposed OLC learns LDR frame representations within specific codebook segments based on exposure bias (short, mid, long) while utilizing the full codebook for HDR priors, enhancing the learning of implicit HDR representations (Fig. 1(b)). This scheme allows the proposed

OLC to represent HDR information by combining LDR representations, similar to the traditional exposure bracket process. The HDR network integrates latent features from the pre-trained VQ decoder and frame context into the fidelity decoder a residual fusing modules, improving HDR image quality. To address frame misalignment, we introduce a parallel alignment module and a dynamic frame merging module to combine LDR frame context with valid regional features. These components collectively enhance the HDR reconstruction process. Experimental results demonstrate that our method outperforms previous methods across various datasets and metrics.

Our contributions can be summarized as follows:

- We introduce an Overlapped Codebook (OLC) scheme for implicitly capturing HDR representations via the VQGAN framework. The OLC aligns with the common exposure bracketing process, achieving improved representation learning ability for multi-exposure HDR imaging.
- We present a dual-decoder HDR network, integrating learned HDR representations from a pre-trained VQ decoder and OLC into the fidelity decoder for high-quality HDR image generation. Additionally, we introduce a parallel alignment module and a frame-selective merging module to address misalignment and incorporate frame context effectively.
- Extensive experiments demonstrate that our HDR network with learned representation in pre-trained OLC achieves superior performance on various datasets and metrics.

2 Related Works

2.1 Multi-exposure HDR Imaging

Multi-exposure HDR imaging generally produces higher-quality results compared to single-image HDR imaging. This is because it can leverage more information from multiple LDR frames. However, taking multiple LDR images can cause hand or object motions, and some LDR images may have under-/over-exposed regions due to scene conditions and exposure biases. Therefore, aligning LDR frames and compensating for saturated areas are the primary concerns in multi-exposure HDR imaging schemes.

Earlier methods proposed a pixel rejection approach for multi-exposure HDR imaging, assuming the images are globally registered. For instance, Grosch [1] uses the color difference of input images as an error map. Jacobs *et al.* [2] measure weighed variance for detecting ghost regions. The registration-based methods were also proposed, which search for similar regions. Kang *et al.* [3] utilize exposure bias information to transform LDR images to the luminance domain and apply optical flow for finding corresponding pixels from non-reference LDR frames. Sen *et al.* [6] introduced a patch-based energy minimization method for jointly optimizing input alignment and HDR image reconstruction.

Recently, CNN-based methods have shown superior performance in various image restoration areas, including HDR imaging. Kalantari *et al.* [7] first proposed a CNN-based method for multi-exposure HDR imaging. They adopted optical-flow estimation for aligning LDR frames in the pre-processing stage, then merged LDR images at the feature level. Wu *et al.* [8] aligned the background through the homography transformation and applied a network with skip-connection for merging. Yan *et al.* [9] proposed a network with a spatial attention module for aligning LDR frames implicitly in the feature domain. Non-local [28] method was also proposed by Yan *et al.*. [17], which constructs a non-local module and triple-pass residual module in the network bottleneck. More recently, Niu *et al.* [11] proposed a GAN-based network for producing a more realistic result, which consists of a generator with reference-based residual merging block. Liu [33] employed a pyramid cascading deformable (PCD) module [34] to align frame features. Vision Transformer (ViT) [12] has also achieved impressive performance in image restoration areas [13,14], and thus applied to HDR imaging. Liu *et al.* [16], Chen *et al.* [35] and Yan *et al.* [32] introduce Transformer-based models for capturing the complex relationship between LDR frames. Further, Song *et al.* [25] proposed a Transformer network with a ghost region detector to make the network focus on valid regions. Tel *et al.* [36] introduced an inter-/intra-frame merging Transformer network with a cross-attention mechanism for utilizing spatial and semantic information.

2.2 Vector Quantization

VQ-VAE [4] was the first to introduce a VQ mechanism to neural networks, which learns discrete code vectors for encoding images. Recently, Esser *et al.* [5] proposed VQGAN for achieving high-quality generated images, which trains the codebook over Transformer architecture and adversarial objectives. The VQ mechanism has also been widely adopted in image restoration areas. Guo *et al.* [29] proposed a super-resolution method with a texture codebook and local autoregressive model for producing finer details. Chen *et al.* [26] introduced a super-resolution network with the pre-trained codebook to leverage learned high-resolution priors. Gu *et al.* [27] proposed a face restoration network that takes advantage of the high-quality feature in the VQ codebook to produce images with realistic face details.

3 Proposed Methods

Given a set of LDR frames with different exposure biases, our target is to compose a single HDR image by the best use of LDR frames' information. Specifically, we propose a 2-step method for multi-exposure HDR imaging which can be summarized as follows:

- **Step 1, Learning implicit HDR representations with the Overlapped Codebook (OLC).**
- **Step 2, HDR reconstruction with the pre-trained OLC and VQ decoder**.

The details of each step are described in the following subsections.

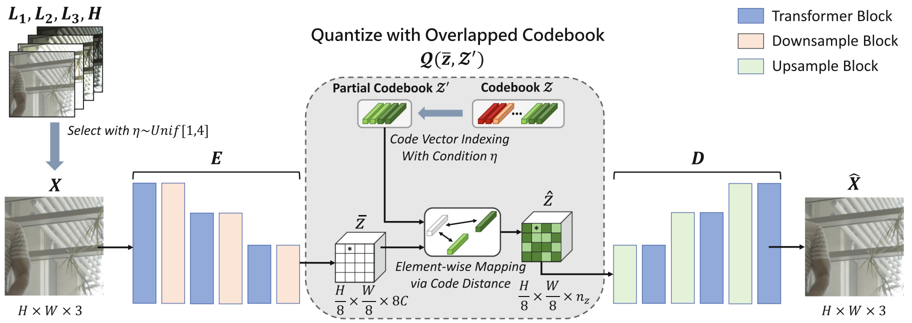


Fig. 2. Illustration of proposed Overlapped Codebook (OLC) scheme with VQGAN framework. In every iteration, we sample $\eta \sim Unif[1,4]$ for randomly selecting input image X and indexing the corresponding codebook segment $\mathcal{Z}'$.

3.1 Learning HDR Representation with the OLC

In this section, we present an OLC, a method that enhances the learning process for capturing HDR representation by aligning with the HDR image generation process. The traditional method for creating ground-truth images in multi-exposure HDR imaging tasks involves merging captured bracketed exposure images [7,30]. For instance, Kalantari *et al.*. [7] employed a triangular weighting function to blend differently exposed static LDR images (S_1, S_2, S_3) as:

$$H = \frac{\sum_i \alpha_i (S_i^{\gamma}/t_i)}{\sum_i \alpha_i}, i = 1, 2, 3, \tag{1}$$

where H is the generated HDR image, γ is a parameter for the gamma-correction function. The α_i is the weights for each LDR frame, which can be defined:

$$\alpha_1 = 1 - \Lambda_1(S_2), \alpha_2 = \Lambda_2(S_2), \alpha_3 = 1 - \Lambda_3(S_2), \tag{2}$$

where $\Lambda_i(\cdot)$ is the triangle function described in Fig. 1(a). To reflect the above-stated weight blending process in multi-exposed LDR fusing, we propose the OLC method that concurrently learns LDR and HDR representations, forming HDR information through a combination of LDR representations. As illustrated in Fig. 1(b), within the OLC framework, each LDR frame is linked to a specific codebook segment based on its exposure bias (short, mid, long) and shares codebook elements with other LDR frames. In contrast, the HDR image is represented using the entire codebook. This distinctive approach employed by OLC improves the capability to represent HDR images within VQ mechanisms.

As illustrated in Fig. 2, we employ the VQGAN framework [5], which consists of encoder E, decoder D, and the overlapped codebook $\mathcal{Z} = \{z_k\}_{k=1}^{K} \in \mathbb{R}^{K \times n_z}$, where K is the codebook size and n_z is the code vector dimension. Given an input image $X \in \mathbb{R}^{H \times W \times 3}$, the encoder produces feature $\bar{z} = E(X) \in \mathbb{R}^{h \times w \times n_z}$. Note that input image X can be each frame of LDR images L_i, $i = 1, 2, 3$ or HDR image H. We randomly select an input from those images with the uniformly sampled parameter $\eta \sim Unif[1,4]$, which can be defined as:

$$X = \begin{cases} L_\eta^\gamma / t_\eta, & \eta \in \{1,2,3\} \\ H, & \eta == 4 \end{cases} \tag{3}$$

where $\gamma = 2.2$ is the parameter of the function and t_η is the exposure bias of the corresponding input LDR image. Note that we use a gamma-correction function on LDR inputs, which maps LDR images into the HDR domain to alleviate the discrepancy between LDR and HDR images. Then, the vector-quantized feature $\hat{z}$ is obtained by finding the nearest neighbors of each feature element in the codebook $\mathcal{Z}$. Different from the common codebook in the VQ scheme, the proposed OLC uses a specific part of codebook $\mathcal{Z}$ following the type of input image X. For instance, when input image X is one of LDR image frame L_i, partial codebook $\mathcal{Z}^i \in \mathbb{R}^{(K/2)\times n_z}$ can be defined as:

$$\mathcal{Z}^i = \{z_{i\times\alpha+1}, z_{i\times\alpha+2}, ..., z_{(i+1)\times\alpha}\}, i \in \{1,2,3\}, \tag{4}$$

where $\alpha = \frac{K}{4}$ is the offset parameter, and i is the index of the LDR frame. When the input image X is an HDR image H, all K code vectors are used ($\mathcal{Z}$). Note that the codebook $\mathcal{Z}^i$ for each LDR frame shares $\frac{K}{4}$ of code vectors. For instance, in the case of partial codebook $\mathcal{Z}^1$, $\mathcal{Z}^2$ for L_1, L_2, they share code vectors $\{z_{\alpha+1}, z_{\alpha+2}, ..., z_{2\times\alpha}\} \in \mathbb{R}^{\alpha\times n_z}$. The VQ process for encoded feature $\bar{z} = E(X)$ can be formulated as:

$$\hat{z}_j = \mathcal{Q}(\bar{z}_j, \mathcal{Z}') = \underset{z_k \in \mathcal{Z}'}{\arg\min} \|\bar{z}_j - z_k\|, \eta \in \{1,2,3,4\}, \\ \text{where } \mathcal{Z}' = \begin{cases} \mathcal{Z}^\eta, & \eta \in \{1,2,3\} \\ \mathcal{Z}, & \eta == 4 \end{cases} \tag{5}$$

where $\mathcal{Q}(\cdot)$ is a quantization function conditioned by the partial codebook $\mathcal{Z}'$, $\hat{z} \in \mathbb{R}^{h\times w\times n_z}$ is a quantized feature, and $j \in \{1,2,...,h\times w\}$. Then, the decoder D reconstructs the result $\hat{X} \approx X$, which can be formulated as:

$$\hat{X} = D(\mathcal{Q}(E(X), \mathcal{Z}')) \in \mathbb{R}^{H\times W\times 3}. \tag{6}$$

Since the quantization function $\mathcal{Q}(\cdot)$ is non-differentiable, we follow previous works [4,5] for backpropagation, which simply copies the gradients from the decoder D to the encoder E. Thus, the codebook, encoder, and decoder can be optimized with loss function $\mathcal{L}_{vq}$, $\mathcal{L}_{rec}$, and $\mathcal{L}_{per}$, which can be defined as:

$$\mathcal{L}_{vq} = \|\text{sg}[E(X)] - \hat{z}\|_2^2 + \beta\|\text{sg}[\hat{z}] - E(X)\|_2^2, \tag{7}$$

where $\beta = 0.25$ is the commitment weight and sg[·] is the stop-gradient operation. It is worth noting that our partial codebook $\mathcal{Z}'$ uses a specific part of the codebook $\mathcal{Z}$ by indexing code vectors. Thus, updating $\mathcal{Z}'$ with Eq. 7 is the

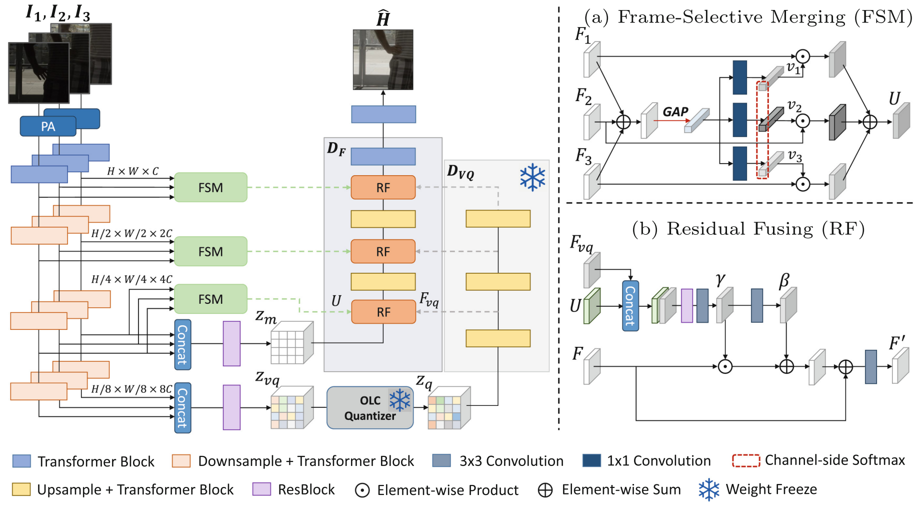


Fig. 3. Illustration of proposed dual-decoder HDR network with fidelity decoder D_F and pre-trained VQ decoder D_{VQ}. The HDR network consists of (a) a Frame-Selective Merging (FSM) unit and (b) a Residual Fusing (RF) unit.

same as updating corresponding code vectors in the master codebook $\mathcal{Z}$. The reconstruction loss and perceptual loss are defined as follows:

$$\mathcal{L}_{rec} = \|\tau(X) - \tau(\hat{X})\|_1, \mathcal{L}_{per} = \|\phi(\tau(X)) - \phi(\tau(\hat{X}))\|_1, \tag{8}$$

where $\tau(\cdot)$ is a μ-law tone-mapping function, and $\phi(\cdot)$ is the pre-trained VGG-16 [20] network. Note that we follow [7,9,11] to train networks more effectively, which apply the tone-mapping function $\tau(\cdot)$ to an HDR image in the training objective. Given an HDR image H, the $\tau(\cdot)$ is defined as follows:

$$\tau(H) = \frac{\log(1 + \mu H)}{\log(1 + \mu)}, \tag{9}$$

where $\mu = 5000$ is a parameter of the tone-mapping function. The final loss for training our VQGAN with the OLC is a weighted sum of all losses:

$$\mathcal{L}_{OLC} = \lambda_{rec}\mathcal{L}_{rec} + \lambda_{per}\mathcal{L}_{per} + \lambda_{vq}\mathcal{L}_{vq} + \lambda_{adv}\mathcal{L}_{adv}, \tag{10}$$

where $\mathcal{L}_{adv} = -\mathbb{E}_{\hat{X}}[D(\hat{X}))]$ is the adversarial loss from discriminator D. With the above codebook structure and learning method, OLC is capable of learning the HDR representations over the LDR subspace.

3.2 HDR Imaging with Learned Representation

Following the acquisition of HDR representation through OLC, we introduce an HDR network designed to generate HDR images from multiple LDR images.

Specifically, we utilize the acquired HDR representations to enhance the realism of HDR images. To achieve this, we employ a pre-trained codebook and VQ decoder, which is introduced in Sect. 3.1. The learned HDR representation proves beneficial in the HDR reconstruction process by compensating for saturated regions and recovering fine details. However, GAN-based methods often encounter fidelity distortions despite improving perceptual quality which is crucial in multi-exposure HDR imaging. Hence, we propose a network with a dual-decoder structure to address both saturated regions and missing details while preserving image fidelity. Given a set of LDR images $L_i \in \mathbb{R}^{H\times W\times 3}, i = 1, 2, 3$, we follow previous works that also use corresponding HDR-mapped images as input $I_i \in \mathbb{R}^{H\times W\times 6}$ for the network:

$$I_i = \left[L_i, L_i^{\gamma}/t_i\right], i = 1, 2, 3, \tag{11}$$

where $\gamma = 2.2$ is the parameter of the gamma-correction function and t_i is the exposure bias (time) of the corresponding LDR frame. We apply a convolution layer to all frames to map them into feature space as: $F_i = \mathrm{Conv}(I_i), i = 1, 2, 3$. Since input LDR frames are not aligned, we construct the parallel alignment (PA) unit at the initial layer in the HDR network for feature-level alignment.

Parallel Alignment. As shown in Fig. 4, the PA module aligns non-reference frames (I_1, I_3) to the reference frame I_2 in the feature space. Features of both frames are concatenated and processed through an offset module with feature-selective mechanisms and multiple receptive fields. Specifically, 3×3 and 5×5 convolutions are applied to generate an offset feature F_o, enabling the PA to handle diverse motion differences. Using the offset feature, the PA aligns the non-reference frame feature F_{NR} with deformable convolution and spatial attention. The aligned input features F_d and F_s are then concatenated to produce the final aligned output F'_{NR}. This parallel approach with dual alignment methods ensures more accurate alignment. This can be defined as:

$$\begin{aligned} F_d &= DF(F_{NR}, \mathrm{Conv}(F_o)), \\ F_s &= SA(F_{NR}, \mathrm{Conv}(F_o)), \\ F'_{NR} &= \mathrm{Conv}([F_d, F_s]), \end{aligned} \tag{12}$$

where $DF(\cdot)$ and $SA(\cdot)$ denote deformable convolution and spatial attention operation, respectively. Note that we have two non-reference frames I_1, I_3, we define two PA for each non-reference frame, $F'_i = PA_i(F_i, F_2), i = 1, 3$. And a convolutional layer applied to reference frame F_2 as: $F'_2 = \mathrm{Conv}(F_2)$.

Following the alignment of non-reference frame features, we establish individual multi-scale encoders to extract features from each LDR frame. Each encoder processes the frame feature $F'_i \in \mathbb{R}^{H\times W\times C}$ and progressively reduces the spatial size to $\frac{H}{8} \times \frac{W}{8} \times 8C$. As depicted in Fig. 3, we combine frame features at both $\frac{H}{4} \times \frac{W}{4}$ and $\frac{H}{8} \times \frac{W}{8}$ scales for the fidelity decoder D_F and pre-trained VQ decoder D_{VQ}, respectively. Given that the pre-trained VQ decoder is trained on $\frac{H}{8} \times \frac{W}{8}$ spatial size, we input the same spatial size of the quantized merged

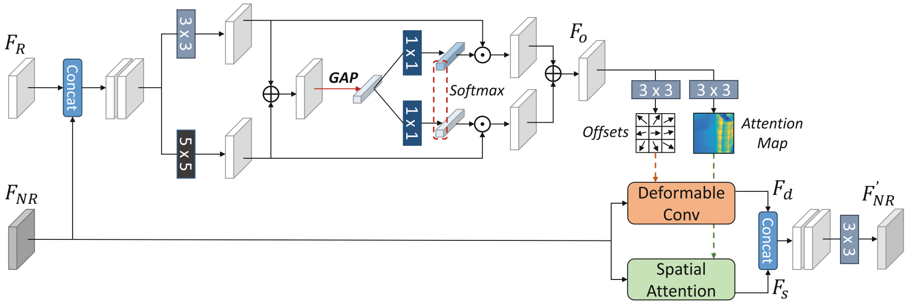


Fig. 4. Illustration of the Parallel Alignment (PA) unit.

Table 1. Quantitative comparison on Kalantari *et al.* [7] and Hu *et al.* [24] dataset. The boldface and underlined numbers denote the best and second-best performances. H.V-2 is HDR-VDP-2 metric. † indicates that the method is excluded from several metrics and experiments since its implementation is not available.

Dataset	Method	PSNR-μ	PSNR-ℓ	PSNR-PU	SSIM-μ	SSIM-ℓ	SSIM-PU	H.V-2
Kalantari [7]	Sen [6]	40.95	38.30	34.44	0.9829	0.9745	0.9783	59.38
	Kalantari [7]	42.74	41.23	36.35	0.9888	0.9846	0.9843	64.42
	DeepHDR [8]	41.91	40.36	35.52	0.9770	0.9602	0.9805	64.78
	AHDRNet [9]	43.70	41.17	37.37	0.9904	0.9856	0.9869	65.11
	NHDRRNet† [17]	42.41	–	–	0.9887	–	–	–
	HDR-GAN [11]	43.92	41.57	37.47	0.9905	0.9865	0.9870	65.58
	ADNet [33]	43.97	41.78	37.62	0.9905	0.9882	0.9867	65.84
	TransHDR† [25]	44.10	41.70	–	0.9909	0.9872	–	–
	CA-ViT [16]	44.32	42.18	37.73	0.9916	0.9884	0.9878	66.33
	HFT† [35]	44.45	42.14	–	0.9920	0.9880	–	66.32
	SCTNet [36]	44.47	42.33	37.95	0.9922	0.9885	0.9887	66.40
	HyHDRNet† [32]	44.64	42.47	–	0.9915	0.9894	–	66.03
	Ours	**44.89**	**42.60**	**38.32**	**0.9935**	**0.9898**	**0.9899**	**66.69**
Hu [24]	Sen [6]	31.51	33.45	30.81	0.9533	0.9630	0.9783	59.38
	Kalantari [7]	42.74	41.23	36.35	0.9888	0.9846	0.9843	63.72
	DeepHDR [8]	41.88	41.96	35.81	0.9790	0.9856	0.9860	63.15
	AHDRNet [9]	46.87	50.70	41.26	0.9959	0.9983	0.9956	64.29
	HDR-GAN [11]	46.69	50.42	41.02	0.9958	0.9988	0.9954	64.33
	ADNet [33]	47.27	51.83	41.44	0.9961	0.9988	0.9957	64.47
	CA-ViT [16]	47.98	52.12	41.68	0.9967	0.9990	0.9960	64.67
	SCTNet [36]	48.18	52.15	41.72	0.9967	0.9991	0.9962	64.84
	HyHDRNet† [32]	48.46	51.91	–	0.9959	0.9991	–	–
	Ours	**48.73**	**52.39**	**42.47**	**0.9970**	**0.9992**	**0.9966**	**65.12**

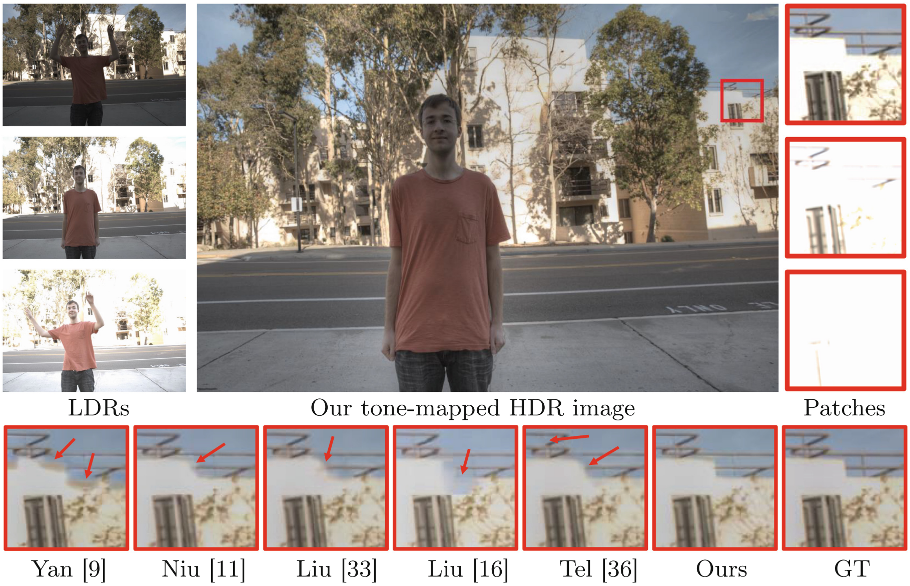
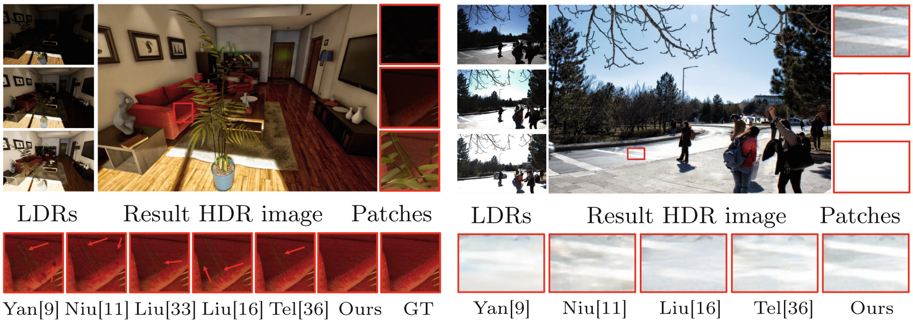


Fig. 5. Visual comparison on a test sample in Kalantari's [7] dataset.

feature $z_q = \mathcal{Q}(z_{vq}, \mathcal{Z}) \in \mathbb{R}^{\frac{H}{8} \times \frac{W}{8} \times 8C}$ into the VQ decoder to minimize discrepancies. Note that we use full codebook $\mathcal{Z}$ to quantize since we target reconstructing HDR images in the HDR network. Conversely, for the fidelity decoder, we input merged features $z_m \in \mathbb{R}^{\frac{H}{4} \times \frac{W}{4} \times 4C}$ with a less reduced scale to preserve structural information. Specifically, the fidelity decoder incorporates features from the VQ decoder and a frame context feature from the encoding stage. Different from existing methods that solely deliver the reference frame feature with a skip connection, we introduce a Frame-Selective Merging (FSM) unit, which aggregates encoded frame contexts for delivering richer frame information to the decoder.

Frame-Selective Merging. In Fig. 3(a), we illustrate the Frame-Selective Merging (FSM) unit. Inspired by [31], FSM employs attention-based mechanisms to aggregate frame features F_i. It first combines input features through summation, then applies global average pooling and a 1×1 convolution to generate a feature vector v. This vector undergoes three individual 1×1 convolutions and channel-wise softmax to produce attention vectors v_i for each frame. The attention vectors v_i are then multiplied by their corresponding frame features, and the processed features are summed to produce the merged context $U = \sum_i (F_i \odot v_i)$. By selecting valid features from each frame, FSM effectively merges frame context, thereby supporting the decoding process.

Residual Fusing. As we stated earlier, our HDR network features a dual-decoder structure. We use a pre-trained VQ decoder D_{VQ} with OLC and add a



(a) Test sample in Hu's dataset.

(b) Test sample in Tursun's dataset.

Fig. 6. Visual comparison on test samples in (a) Hu's dataset and (b) Tursun's dataset. Note that samples in Tursun's dataset has no ground-truth HDR images.

fidelity decoder D_F for HDR reconstruction. To leverage the VQ decoder's HDR representation capabilities, we propose a Residual Fusing (RF) module. As shown in Fig. 3(b), RF takes intermediate features F_{vq} from D_{VQ} and merged contexts U from FSM to fuse internal features in D_F. Both F_{vq} and U are concatenated and fed into a resblock to produce parameter features γ and β. RF then fuses the input feature with γ and β through affine transformation, finally producing output feature F' with a residual connection. This can be defined as:

$$\begin{aligned} \gamma, \beta &= \mathrm{Conv}([U, F_{vq}]), \\ F' &= (\gamma \odot F + \beta) + F. \end{aligned} \tag{13}$$

With this residual fusing method, RF is able to incorporate VQ features and context while retaining image fidelity with the residual connection.

The training objective of our HDR network is the combination of three losses: 1) reconstruction loss $\mathcal{L}_{rec}$ for maintaining data fidelity; 2) perceptual loss $\mathcal{L}_{per}$ for producing realistic details; 3) mapping loss $\mathcal{L}_{map}$ for mapping extracted features to code vectors in the learned codebook. Given the ground-truth HDR image H and a predicted HDR image $\hat{H}$, the $\mathcal{L}_{rec}, \mathcal{L}_{per}$ can be defined as:

$$\mathcal{L}_{rec} = \|\tau(H) - \tau(\hat{H})\|_1, \mathcal{L}_{per} = \|\phi(\tau(H)) - \phi(\tau(\hat{H}))\|_1, \tag{14}$$

where $\phi(\cdot)$ is pre-trained VGG-16 network [20]. The mapping loss $\mathcal{L}_{map}$ calculates the distance between the extracted feature $z_{gt} \in \mathbb{R}^{\frac{H}{8} \times \frac{W}{8} \times 8C}$ in the HDR network and ground-truth VQ representation $z_{gt} = \mathcal{Q}(E(H), \mathcal{Z})$, defined as:

$$\mathcal{L}_{map} = \|z_{vq} - z_{gt}\|_2^2. \tag{15}$$

The final loss $\mathcal{L}_{HDR}$ is weighted sum of all losses:

$$\mathcal{L}_{HDR} = \mathcal{L}_{rec} + \lambda_{per}\mathcal{L}_{per} + \lambda_{map}\mathcal{L}_{map}. \tag{16}$$

4 Experiments

4.1 Dataset and Metrics

Dataset. We train and test our method on Kalantari *et al.*'s dataset [7] and Hu *et al.*'s dataset [24]. Specifically, Kalantari *et al.*'s dataset consists of 74 samples for training and 15 samples for testing. Each data pair contains three LDR images that are captured with $\{-2, 0, +2\}$ or $\{-3, 0, +3\}$ of exposure bias sets and a single HDR image. Hu *et al.*'s dataset [24] synthesized with the game engine, and captured with an exposure bias of $\{-2, 0, +2\}$.

Evaluation Metrics. We compute metrics on both results linear HDR image $\hat{H}$ and tone-mapped HDR image $\tau(\hat{H})$. The PSNR-ℓ, SSIM-ℓ are calculated between linear HDR image H, $\hat{H}$ and PSNR-μ, SSIM-μ are calculated between tone-mapped images $\tau(H), \tau(\hat{H})$. Furthermore, we also measure HDR-VDP-2 [18], which evaluates the quantitative quality of HDR images on specified display and luminance conditions. Lastly, we report the PU21 [19] metric, which measure the similarity between perceptually uniform values of the HDR images.

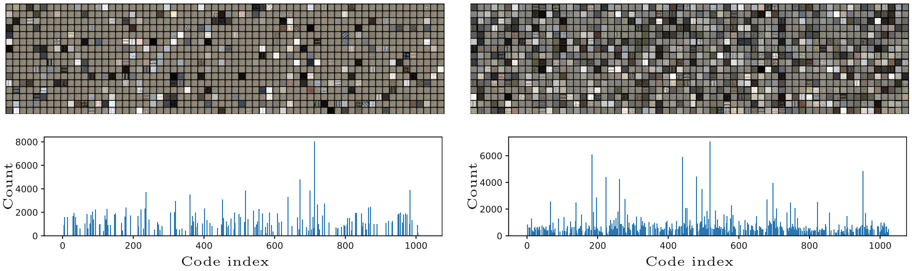
4.2 Training Details

For training both the OLC and the HDR network, we crop patches of size 256×256 with a stride of 64 from training samples. Further, we also apply a set of augmentation, including horizon/vertical flipping and rotation. All experiments are implemented with the Pytorch framework and a single NVIDIA RTX 3090 Ti GPU. We adopt Adam optimizer [15] with 1e-4 learning rate for training generators in OLC and HDR network. For the discriminator in VQGAN, a learning rate of 4e-4 is set. The number of code vectors in the OLC is set as $K = 1024$ and the base channel size of the HDR network is $C = 32$.

4.3 Comparison with Previous Methods

Quantitative Comparison. Table 1 shows a quantitative comparison with previous methods on Kalantari's dataset [7] and Hu's dataset [24]. Generally, deep learning-based methods [7–9, 11, 33] show improved performance compared to patch-based [6, 21] algorithms. Furthermore, Transformer-based methods [16, 32, 35, 36] outperform previous methods by notable margins. Our method achieves the best performance on most metrics, including HDR-VDP-2 and PU21. This result implies our method is not only producing more realistic HDR images but also robust on certain display and luminance conditions.

Qualitative Comparison. We further evaluate the qualitative results in Fig. 5 and Fig. 6. Note that we use a tone-mapping function of Photomatix to visualize HDR images. Figure 5 displays the ability to reconstruct heavily saturated regions. AHDRNet [9], ADNet [33], and HDR-GAN [11] produce blurry detail



(a) Vanilla VQ Codebook (b) Overlapped Codebook

Fig. 7. Code vector visualization (first row) and distribution (second row) in the vanilla VQ codebook and proposed Overlapped codebook (OLC).

Table 2. Performance on Test samples in [7] with vanilla codebook and OLC. K denotes the number of code vectors.

Method	PSNR-μ	PSNR-ℓ	H.V-2
Vanilla (K=512)	44.38	42.20	66.31
OLC (K=512)	44.55	42.36	66.42
Vanilla (K=1024)	44.57	42.32	66.44
OLC (K=1024)	44.89	42.60	66.69

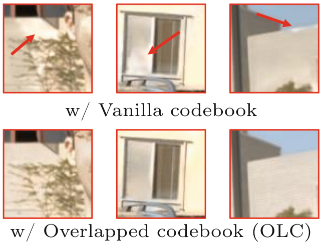


Fig. 8. Visual comparison on vanilla codebook and OLC (K=1024).

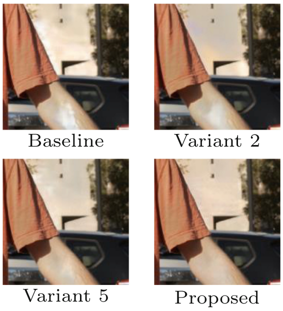
component and edges regions. CA-ViT [16] and SCTNet [36] show the resulting image with better-detailed regions, but there are distorted region remains on the edges. In contrast, our method produces clear edges and fine details without distortion. In Fig. 6 (a), a large motion difference exists between LDR frames. Different from other methods that leave ghosting artifacts on moving objects, our method effectively address misalignment with PA modules and produces result HDR images without undesired artifacts. We also compare our method on the Tursun *et al.* [23] dataset, which has no ground-truth HDR image in Fig. 6 (b). Since the scene information in the reference frame and high exposure frame was severely lost due to over-exposure, other methods failed to compensate for saturated regions from valid regions in other frames. In contrast, our method shows more realistic HDR images in extreme cases. We report additional quantitative and qualitative results in the supplementary materials.

4.4 Analysis on the Proposed OLC

As previously discussed, proposed OLC significantly enhances the capacity to learn implicit HDR representations. In Fig. 7, we provide visualizations of code vectors within the pre-trained VQGAN framework and display the code index distribution for reconstructing HDR images. It's important to note that both

Table 3. Ablation on proposed components. *Sum* and *Concat* in variants 3, 4 denote the frame merging method.

Method	PSNR-μ	PSNR-ℓ	H.V-2
1. Baseline	43.92	41.77	65.79
2. + PA	44.20	41.94	66.02
3. + PA + *Sum*	44.31	42.11	66.15
4. + PA + *Concat*	44.38	42.22	66.22
5. + PA + FMU	44.49	42.30	66.35
6. + PA + FMU + D_{VQ}	44.74	42.51	66.60
7. + PA + FMU + D_{VQ} + RF	44.89	42.60	66.69


Baseline Variant 2

Variant 5 Proposed

Fig. 9. Visual comparison on variants in ablation.

the vanilla VQ codebook (a) and the OLC (b) are trained under identical conditions, including training iteration and network settings. The visualization illustrates that the proposed OLC explores a more diverse range of HDR representations, learning additional valid code vectors and utilizing them to restore HDR images. Furthermore, we compare the performance of OLC with the vanilla codebook in Table 2. OLC demonstrates superior performance in reconstructing HDR images, particularly with a larger codebook size (K). In Fig. 8, we showcase predicted HDR patches with the vanilla codebook (first row) and OLC (second row). Compared to the vanilla codebook, OLC exhibits enhanced capability in restoring saturated and detailed regions. These results affirm that our OLC offers improved representation learning ability, consequently enhancing performance without additional computational burden in reconstructing HDR images.

4.5 Impact of Proposed Modules

In Table 3 and Fig. 9, we conduct an ablation study on Kalantari's dataset to demonstrate the effectiveness of the proposed modules in the HDR network. The Baseline model consists of an encoder and fidelity decoder. Variants 3 and 4 merge frame contexts by summing ($U = F_1 + F_2 + F_3$) or concatenating ($U = \text{Conv}([F_1, F_2, F_3])$) instead of using the FMU. Variants 3–6 also incorporate merged context U or VQ feature F_{vq} without the RF module. Variant 2, with PA modules, reduces ghosting artifacts and improves quality in misalignment regions. Variant 5, with FMU modules, better compensates for saturated regions. Variant 7, the proposed network that incorporating all proposed components, achieves the best performance, producing more realistic HDR images. These results validate the effectiveness of each proposed module and the pre-trained VQ component in enhancing HDR reconstruction performance.

5 Conclusion

We proposed an Overlapped Codebook (OLC) scheme for multi-exposure HDR imaging, which effectively learns implicit HDR representations within the

VQGAN framework by modeling the HDR generation process in exposure bracketing. Additionally, we introduced a dual-decoder HDR network that leverages these acquired HDR representations from the pre-trained OLC to produce high-quality HDR images. Our network includes a parallel alignment module to correct misalignment among LDR frames and features frame-selective merging and residual fusing modules to integrate HDR representations with valid frame contexts during decoding. Extensive experiments demonstrate significant improvements with our method on benchmark datasets.

Acknowledgement. This work was supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government(MSIT) [NO.RS-2021-II211343, Artificial Intelligence Graduate School Program (Seoul National University)] [No.RS-2021-II212068, Artificial Intelligence Innovation Hub], and in part by Samsung Electronics Co., Ltd.

References

1. Grosch, T.: Fast and robust high dynamic range image generation with camera and object movement. Vis. Model. Vis. RWTH Aachen **277284**, 2 (2006)
2. Jacobs, K., Loscos, C., Ward, G.: Automatic high-dynamic range image generation for dynamic scenes. IEEE Comput. Graph. Appl. **28**, 84–93 (2008)
3. Kang, S., Uyttendaele, M., Winder, S., Szeliski, R.: High dynamic range video. ACM Trans. Graph. (TOG) **22**, 319–325 (2003)
4. Van Den Oord, A., Vinyals, O.: Neural discrete representation learning. In: Advances In Neural Information Processing Systems. vol. 30 (2017)
5. 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)
6. Sen, P., Kalantari, N., Yaesoubi, M., Darabi, S., Goldman, D., Shechtman, E.: Robust patch-based HDR reconstruction of dynamic scenes. ACM Trans. Graph. **31**(6), 1–11 (2012)
7. Kalantari, N., Ramamoorthi, R.: Deep high dynamic range imaging of dynamic scenes. ACM Trans. Graph. **36**, 1–12 (2017)
8. Wu, S., Xu, J., Tai, Y., Tang, C.: Deep high dynamic range imaging with large foreground motions. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 117–132 (2018)
9. Yan, Q., et al.: Attention-guided network for ghost-free high dynamic range imaging. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 1751–1760 (2019)
10. Goodfellow, I., et al.: Generative adversarial networks. Commun. ACM **63**, 139–144 (2020)
11. Niu, Y., Wu, J., Liu, W., Guo, W., Lau, R.: HDR-GAN: HDR image reconstruction from multi-exposed LDR images with large motions. IEEE Trans. Image Process. **30**, 3885–3896 (2021)
12. Dosovitskiy, A., et al.: An image is worth 16×16 words: transformers for image recognition at scale. In: International Conference on Learning Representations (ICLR) (2021)

13. Liu, Z., et al.: Swin transformer: hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 10012–10022 (2021)
14. Zamir, S., Arora, A., Khan, S., Hayat, M., Khan, F., Yang, M.: 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)
15. Kingma, D., Ba, J.: Adam: a method for stochastic optimization. In: International Conference on Learning Representations (ICLR) (2015)
16. Liu, Z., Wang, Y., Zeng, B., Liu, S.: Ghost-free high dynamic range imaging with context-aware transformer. In: European Conference on Computer Vision, pp. 344–360 (2022)
17. Yan, Q., et al.: Deep HDR imaging via a non-local network. IEEE Trans. Image Process. **29**, 4308–4322 (2020)
18. Mantiuk, R., Kim, K., Rempel, A., Heidrich, W.: HDR-VDP-2: a calibrated visual metric for visibility and quality predictions in all luminance conditions. ACM Trans. Graph. (TOG) **30**, 1–14 (2011)
19. Azimi, M.: PU21: a novel perceptually uniform encoding for adapting existing quality metrics for HDR. In: 2021 Picture Coding Symposium (PCS), pp. 1–5 (2021)
20. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. In: International Conference on Learning Representations (ICLR) (2015)
21. Hu, J., Gallo, O., Pulli, K., Sun, X.: HDR deghosting: how to deal with saturation?. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1163–1170 (2013)
22. Zimmer, H., Bruhn, A., Weickert, J.: Freehand HDR imaging of moving scenes with simultaneous resolution enhancement. Comput. Graph. Forum. **30**, 405–414 (2011)
23. Tursun, O., Akyüz, A., Erdem, A., Erdem, E.: An objective deghosting quality metric for HDR images. Comput. Graph. Forum. **35**, 139–152 (2016)
24. Hu, J., et al.: Sensor-realistic synthetic data engine for multi-frame high dynamic range photography. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, pp. 516–517 (2020)
25. Song, J., Park, Y., Kong, K., Kwak, J., Kang, S.: Selective transHDR: transformer-based selective HDR imaging using ghost region mask. In: European Conference on Computer Vision, pp. 288–304 (2022)
26. Chen, C., et al.: Real-world blind super-resolution via feature matching with implicit high-resolution priors. In: Proceedings of the 30th ACM International Conference on Multimedia, pp. 1329–1338 (2022)
27. Gu, Y., et al.: VQFR: blind face restoration with vector-quantized dictionary and parallel decoder. In: European Conference on Computer Vision, pp. 126–143 (2022)
28. Wang, X., Girshick, R., Gupta, A., He, K.: Non-local neural networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 7794–7803 (2018)
29. Guo, B., Zhang, X., Wu, H., Wang, Y., Zhang, Y., Wang, Y.: LAR-SR: a local autoregressive model for image super-resolution. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 1909–1918 (2022)
30. Debevec, P., Malik, J.: Recovering high dynamic range radiance maps from photographs. Seminal Graph. Pap. Pushing Boundaries **2**, 643–652 (2023)

31. 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)
32. Yan, Q., Chen, W., Zhang, S., Zhu, Y., Sun, J., Zhang, Y.: A unified HDR imaging method with pixel and patch level. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 22211–22220 (2023)
33. Liu, Z., et al.: ADNet: attention-guided deformable convolutional network for high dynamic range imaging. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 463–470 (2021)
34. Wang, X., Chan, K., Yu, K., Dong, C., Change Loy, C.: EDVR: video restoration with enhanced deformable convolutional networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (2019)
35. Chen, R., et al.: Improving dynamic HDR imaging with fusion transformer. Proc. AAAI Conf. Artif. Intell. **37**, 340–349 (2023)
36. Tel, S., et al.: Alignment-free HDR deghosting with semantics consistent transformer. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 12836–12845 (2023)

Deformable Multi-Scale Network for Snow Removal in Video

Runlin He, Gang Zhou(✉), Tianhao Xue, Zhaoxi Liu, and Zhenhong Jia

Key Laboratory of Signal Detection and Processing, Department of Computer Science and Technology, Xinjiang University, Urumqi, China
gangzhou_xju@126.com

Abstract. Snowfall severely degrades outdoor video visibility while reducing the performance of subsequent vision tasks. Although video recovery methods based on deep learning have achieved amazing accomplishments, video snow removal still faces problems such as varying scales and intricate trajectories of snowflakes, which makes it difficult to remove snowflakes and easy to create artifacts on moving objects. To address these issues, we propose a deformable multi-scale video desnowing network. Specifically, we design a multi-scale pseudo-3D residual block(MSRB-P3D) that can effectively remove snowflakes of different scales. Furthermore, a deformable large kernel attention 3Dblock(D-LKA 3Dblock) is introduced to capture the inter-frame dynamic information and reduce the artifacts. Due to the scarcity of dataset, we proposed a new dataset named Synthetic and Real Snowy Video Dataset(SRSVD). Extensive experiments have proven that our proposed method not only outperforms other state-of-the-art methods on both synthetic and real snowy videos, but also effectively improves performance on subsequent vision task.

Keywords: Video desnowing · Multi-scale pseudo-3D residual block · Deformable large kernel attention 3Dblock

1 Introduction

In the past decade, computer vision technology has been a research hotspot in the field of deep learning. With the continuous deepening and maturing of research, it is widely used in many aspects such as automatic driving and video surveillance. However, inclement weather (e.g., rain, snow, etc.) often affects the visual quality of the images and videos, thereby degrading the performance of subsequent vision tasks such as object tracking. As a result, video recovery in adverse weather has become an issue of great concern.

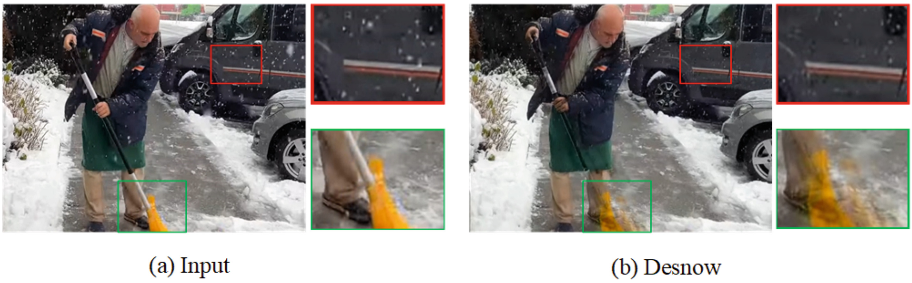
G. Zhou—This work was supported by National Natural Science Foundation of China (No.62166040, No.62261053) and Tianshan Talent Training Project - Xinjiang Science and Technology Innovation Team Program (2023TSYCTD0012).

A. Antonacopoulos et al. (Eds.): ICPR 2024, LNCS 15332, pp. 287–301, 2025.
https://doi.org/10.1007/978-3-031-78125-4_20

Snowfall is a very common natural phenomena that severely degrades outdoor video visibility while reducing the performance of subsequent vision tasks. Due to the lack of video snow datasets, research on video desnowing is relatively rare. Some scholars [1,2] believed that snow and rain share similar characteristics. They categorized the rain and snow removal into the same processing task that could be solved with the model-driven methodology. Such conventional methods [1,3–5] utilized physical models of rain/snow to encode a variety of well-designed prior knowledge into an optimization problem. Nevertheless, these approaches refer to deal with specific rainfall/snowfall patterns, but have difficulty handling complex real-world scenarios.

Over the past few years, deep learning based methods have been favored in video restoration, and achieved better performance on rain/snow removal, such as [6–8]. Although [6,7] achieved good results in rainy video, they tended to perform poorly when directly applied to video snow removal tasks. [8] obtained good performance on video desnowing, but it did not take into account the domain gap between synthetic and real-world data.

Taking S2VD [6] as an example, we retrain it with our dataset SRSVD, and test it on real snowy video. As shown in Fig. 1, the snowflakes are hard to removed and artifacts appear on moving object. We analyze that it is due to the differences in physical properties between rain and snow: Snowflakes are more opaque and multi-scale than rain streaks, which make it difficult to remove completely. The motion trajectories of snowflakes are more complex than rain streaks, and are more prone to producing artifacts in the process of video desnowing.


(a) Input

(b) Desnow

Fig. 1. The performance of S2VD [6] on a real snowy video. Red box indicates area where snowflakes have not been completely removed and green box indicates area where artifacts appear.

To address these issues, we propose a deformable multi-scale video desnowing network. In view of the multi-scale characteristic of snowflakes, a multi-scale pseudo-3D residual block (MSRB-P3D) is designed to capture snowflakes of different sizes and shapes. In addition, we introduce a deformable large kernel attention 3Dblock (D-LKA 3Dblock) to handle spatio-temporal features more effectively. Our contributions are summarized as follows:

1. We design MSRB-P3D to extract spatial and temporal information respectively through temporal convolution and spatial convolution module. In this novel module, a dual-channel approach is applied to extracting multi-scale spatial features, so as to improve the network's ability to detect snowflakes of different scales.
2. A D-LKA 3Dblock is introduced, which allows the receptive field to be freely deformed based on features by adjusting the learnable offsets. In this way, our network can better acquire dynamic information between video frames and reduce the artifacts caused by 3D convolution.
3. We propose a new snowy video dataset(SRSVD) and a deformable multi-scale network(DMSNet) for Snow Removal in Video. On this basis, we conduct experimental comparisons with numerous state-of-the-art algorithms and ablation experiments on our method. Furthermore, we evaluate the performance of snow removal methods on object tracking.

2 Related Work

2.1 Video Desnowing Datasets

While there are several single image snow removal datasets available, such as Snow100K [9], CSD [10] and SnowCityScapes [11], video snow removal dataset receives little attention. Chen et al. [8] synthesized the first high-quality video snow removal dataset(RVSD) by Unreal Engine, and it has been publicly available. RVSD includes 110 synthetic snowy videos, in which 80 videos are used for training and 30 videos for testing, and video resolutions range from 480p to 4k. However, the RVSD lacks the real snowy videos.

2.2 Single Image Deraining/Desnowing

Single image rain/snow removal has attracted academic attention earlier. Chen et al. [12] improved the robustness of unsupervised single image deraining using double-contrast learning. Considering the lack of background information, Chen et al. [13] proposed a network called JSTASR and reformulate the snow model to achieve end-to-end learning. In [10], the contradict channel loss and hierarchical decomposition paradigm was proposed to improve snow removal performance. Liu et al. [9] proposed a new subnetwork expansion pyramid that enhances the ability to extract features in scale invariance. Since temporal information is not utilised, these methods have difficulty in achieving satisfactory results on video desnowing.

2.3 Video Deraining/Desnowing

Different from single image, the video has a lot of temporal redundancies that provides more information for video restoration, which can assist in removing rain/snow. Yue et al. [6] modeled the video as a 3D Hidden Markov Model and designed a semi-supervised rain removal algorithm to improve the performance on real-world scenarios. Zhang et al. [7] designed an end-to-end video deraining network called ESTINet to extract advanced spatial features and temporal correlations respectively. Yang et al. [14] first introduced a self-supervised video deraining method that exploits temporal consistency to further improve the quality of rain removal. Wang et al. [15] devised a new video rain synthesis model with the concept of rain streak motions and developed a recurrent disentangled deraining network. Yang et al. [16] proposed ViWS-Net that can recover videos under various adverse weather conditions. Very recently, Chen et al. [8] synthesized the first video snow removal dataset, but they did not consider the domain gap between synthetic and real-world data, which makes it hard to desnow in real-world scenarios. Xue et al. [17] proposed a two-stage desnowing network and solved the domain gap problem by a domain adaptive module. It can be seen that there are fewer studies on video snow removal. As [8] said, deep learning based video desnowing remains an under-researched area.

3 The Proposed Method

3.1 Synthetic and Real Snowy Video Dataset

Unlike previous work [8] that only considered synthetic videos, We propose a new dataset named Synthetic and Real Snowy Video Dataset(SRSVD), which includes both synthetic and real snowy videos. We have collected 62 videos in total, among which the training set includes 24 groups of synthetic snowy videos and 8 groups of real snowy videos(180 frames/group, 30fps), and the test set includes 15 groups of synthetic snowy videos and 15 groups of real snowy videos(60 frames/group, 30fps). All of them are nighttime videos and the resolution is 640*480 pixels. In order to ensure the richness and complexity of snowy videos, our dataset covers a variety of scenes such as pedestrians, buildings, streets, cars, nature scenes, etc.

By downloading from websites and shooting in reality, we collect a lot of real snowy videos and snow-free videos in different scenarios. Then, we employ Adobe After Effect [18] to composite our synthetic snowy videos by overlaying snow layers on snow-free videos. The snow layers are made by both snow video materials and the “CC Snowfall” simulation of Adobe After Effect. Snow video materials are collected from the web, which have black backgrounds and dynamic snowflakes. “CC Snowfall” is used to simulate snow effects with depth of field, light effects and motion blur.

When compositing different snow videos, we change the parameters of CC Snowfall (speed, scene depth, size, flakes, amount, etc.) and deform the snow video materials (time-stretching and spatial deformation, etc.). This is useful for producing more realistic snow layers.

3.2 Overall Architecture

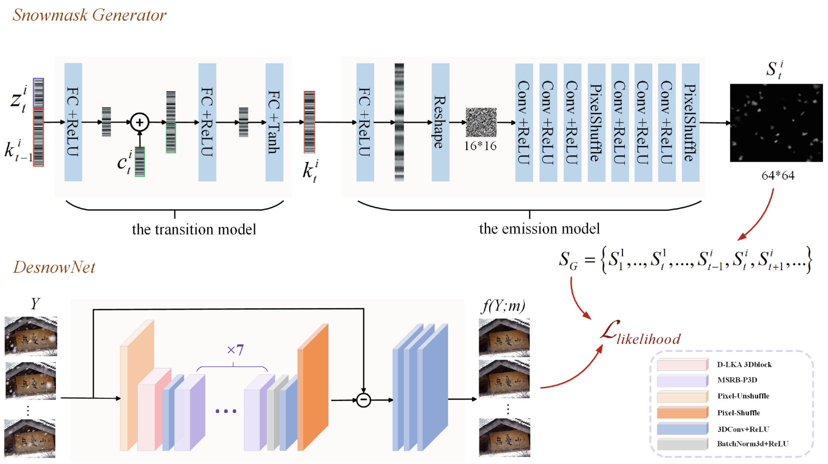


Fig. 2. Overall network architecture of our method.

The overall network is shown in Fig. 2, which consists of DesnowNet and Snowmask Generator. As shown in formula(1), the snowy video Y is decomposed into three parts:

$$Y = f(Y;m) + S + \varepsilon \tag{1}$$

where $f(Y;m)$ and S is the output of DesnowNet and Snowmask Generator respectively, and ε is the residual element that is assumed to follow a zero-mean Gaussian distribution with variance σ^2 at point of pixel. m is the model parameter of DesnowNet.

According to the semi-supervised rain removal algorithm proposed by [6], the snow-free background video is encoded via 3D Markov Random Field(MRF) probability distribution as formula(2):

$$p(W) \propto \exp\{-\rho V\} \tag{2}$$

where ρ is a manual hyper-parameter, and

$$V = \sum_{i,j,t} (\gamma_1 * |f_{i+1,j,t} - f_{ijt}| + \gamma_2 * |f_{i,j+1,t} - f_{ijt}| + \gamma_3 * |f_{i,j,t+1} - f_{ijt}|) \tag{3}$$

f_{ijt} denotes the element of $f(Y;m)$ at location (i,j,t). γ_1, γ_2 and γ_3 are manual hyper-parameters that can be understood as the smoothness constraints on horizontal pixels, vertical pixels, and temporal dimensions respectively.

As for synthetic snowy video, the known groundtruth X can be further embedded as another strong prior as formula(4):

$$p(W) \propto \exp(-I_{prior} - \rho V) \tag{4}$$

where

$$I_{prior} = \frac{\|f(Y;m) - X\|_2 + \sum_{i,j,t} \|(f_{ijt} - f_{ij,t-1}) - (X_{ijt} - X_{ij,t-1})\|_2}{\mu_0^2} \quad (5)$$

μ_0 is a hyper-parameter close to zero. In the same vein as f_{ijt}, X_{ijt} denotes the element of X at location (i, j, t).

In addtion, manual hyper-parameters ρ, σ^2, γ_1, γ_2 and γ_3 and μ_0^2 are set as 0.5, 1, 1, 1, 2 and 1e-6, respectively.

DesnowNet. After multiple consecutive video frames input, DesnowNet reduces required computational resources through pixel-unshuffle. Then, the D-LKA 3Dblock generates adaptive 3D large convolutional kernel in a learnable manner to improve the dynamic ability to capture inter-frame information, which enhances the power to handle background changes and object movements between frames with the help of more free receptive field. In addition, unlike 3D convolution that extracts spatiotemporal features simultaneously, MSRB-P3D extracts temporal and spatial information separately. It adopts dual channel spatial convolution kernels to extract features of different scales, and enables the network to detect snowflakes of different sizes and shapes. Then, the residual connection after pixel-shuffle makes the network to learn and converge better, and finally 3D convolutions are used for refinement to obtain sequences without snow.

Snowmask Generator. The Snowmask Generator consists of a transition model and an emission model. The transition model has three fully connected layers, where k_t^i(features=128) represents the hidden state variable of t-th frame in the i-th snowy video, and c_t^i(features=64) is introduced to account for the variation of snow appearances or patterns. The noise vector z_t^i(features=64) encodes the random factors(e.g., wind, camera motion, etc.) at time t. Firstly, z_t^i and hidden state variables of the previous frame(i.e. k_{t-1}^i) pass through the first fully connected layer(FC). The output of first FC layer concats with c_t^i, and then passes through the last two FC layers to generate the current frame's k_t^i.

After that, k_t^i passes through the emission model to generate the snow mask S_t^i of the current frame. In this process, the FC layer of emission model outputs 256 neurons, which are reshaped to 16*16 in the next step. Finally, the snow mask is obtained through multiple convolutional layers and pixelshuffle operations. For the convenience of formula representation, the collection of all snow masks are defined as S_G.

3.3 Multi-scale Pseudo-3D Residual Block

In order to improve the multi-scale perception ability of the network, we design multi-scale pseudo-3D residual block(MSRB-P3D) that can be seen in Fig. 3.

Compared to 2D convolution, 3D convolution can extract both temporal and spatial features simultaneously. However, the multi-scale design based on 3D convolution will inevitably lead to a significant increase in network parameter count and FLOPs. Therefore, we adopt the idea of pseudo 3D networks, dividing 3D convolution into a spatial convolution module and a temporal convolution, which are connected in series to extract spatial and temporal information respectively. Due to the excellent performance of residual connectivity, the MSRB-P3D has two levels of nested residual learning.

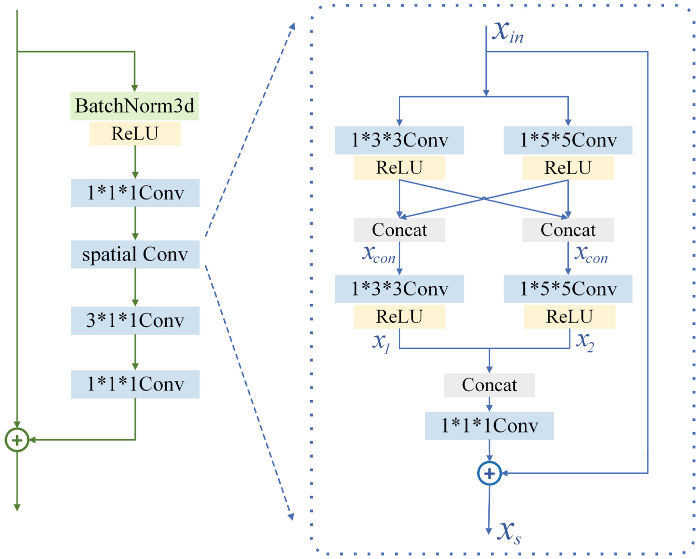


Fig. 3. Multi-scale Pseudo-3D Residual Block.

The spatial convolution module adopts a dual path parallel structure with x_{in} as input. First, we utilize the different pseudo-3D spatial convolution kernels to obtain features of different scales, and concatenate them to generate x_{con}. Then, the multi-scale feature extraction and integration are performed again to generate the features of double branch (i.e. x_1 and x_2). Ultimately, we concatenate them and do residual connection with x_{in} to generate the output x_s. As a result, the multi-scale spatial features x_s are beneficial to improve the network's ability to capture snowflakes of different sizes and shapes.

3.4 Deformable Large Kernel Attention 3Dblock

Combining the broad receptive field of large kernel convolutional attention mechanism with the flexiblity of deformable convolution, Azad et al. [19] proposed D-LKANet that is capable of handling complex visual information and has made significant improvements in the field of medical segmentation.

The ability of small receptive fields to capture inter frame information is insufficient, especially when dealing with moving objects between frames. It is the primary reason for the appearance of artifacts. There is temporal correlation between video frames, which can lead to information leakage during video reconstruction [20]. As a result, the content of the current frame appears on the neighbouring frames. Therefore, we introduce a deformable large kernel attention 3Dblock(D-LKA 3Dblock) that improves the ability to capture the inter-frame dynamics and reduces the artifacts caused by 3D convolution. The D-LKA 3Dblock is shown in Fig. 4, which can be formulated as:

$$output = 3DConv(Attention \otimes F') + F \tag{6}$$

$$F' = GELU(3DConv(F)) \tag{7}$$

where F denotes the input features of this block and F' is the input features of LKA3d-Deform. *Attention* is obtained by 3D convolution, large lernel dilated 3D convolution(kernel size=(7,7, 7), dilation=3) and deformable 3D convolution [21]. Unlike traditional attention methods, this network does not require additional normalization functions (sigmoid or Softmax), and each value represents the relative importance of the corresponding feature. Operator $\otimes$ is element-wise multiplication operation. The residual connection is corresponding to formula(6).

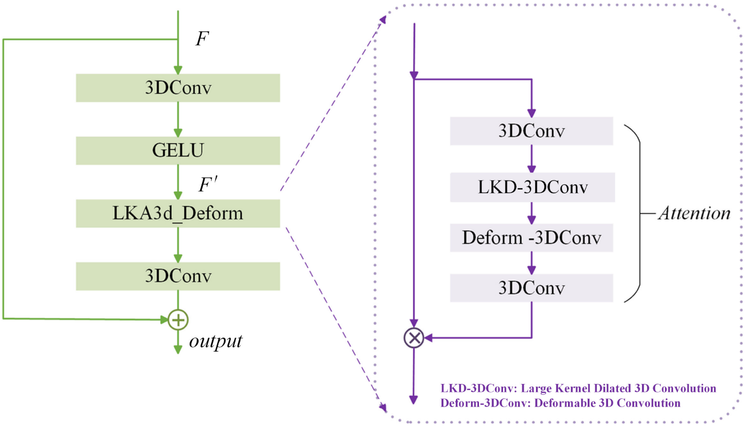

F
3DConv
GELU
F'
LKA3d_Deform
3DConv
output
3DConv
LKD-3DConv
Deform -3DConv
3DConv
Attention
LKD-3DConv: Large Kernel Dilated 3D Convolution
Deform-3DConv: Deformable 3D Convolution

Fig. 4. Deformable Large Kernel Attention 3Dblock.

By this way, a large 3D convolution kernel can be constructed with fewer parameters and computational complexity. In this module, the large kernel provides a receptive field similar to the self attention mechanism. What's more, deformable 3D convolution adjusts the learnable offset to allow the receptive field to deform freely. The deformable convolution can assist with interframe feature extraction to capture changes in background and object movements, and eliminate the video artifacts caused by the fixed field. As a result, the ability of the network to capture the temporal correlation is improved, which is essential for video desnowing.

3.5 Loss Functions

Based on the inference and learning algorithm of [6], the total loss of the network is divided into four parts. As shown in formula(8):

$$\mathcal{L}_{Overall} = \mathcal{L}_{likelihood} + \mathcal{L}_{MRF} + \lambda(\mathcal{L}_{MSE} + \mathcal{L}_{interframe)} \tag{8}$$

where λ is equal to 1 when the input of the DesnowNet is synthetic snowy videos, otherwise $\lambda = 0$(real snowy videos), and

$$\mathcal{L}_{likelihood} = \frac{1}{2\sigma^2} \|Y - f(Y;m) - S_G\| \tag{9}$$

$$\mathcal{L}_{MRF} = \rho V \tag{10}$$

$$\mathcal{L}_{MSE} = \frac{\|f(Y;m) - X\|_2}{\mu_0^2} \tag{11}$$

$$\mathcal{L}_{interframe} = \frac{\sum\limits_{i,j,t} \|(f_{ijt} - f_{ij,t-1}) - (X_{ijt} - X_{ij,t-1})\|_2}{\mu_0^2} \tag{12}$$

The $\mathcal{L}_{likelihood}$ is derived from formula(1). It represents the similarity between snowy video Y and estimation results $f(Y;m) + S_G$, where S_G and $f(Y;m)$ are the output of Snowmask Generator and DesnowNet, respectively; The $\mathcal{L}_{MRF}$ originates from the MRF prior in formula(2); The $\mathcal{L}_{MSE}$ and $\mathcal{L}_{interframe}$ are utilized only for synthetic snowy videos, which correspond to I_{prior} in formula(5). In addition, $\mathcal{L}_{MSE}$ is the inaccuracy between the groundtruth X and desnowing video $f(Y;m)$. $\mathcal{L}_{interframe}$ is time loss based on interframe differences. The hyper-parameter μ_0^2 can help adjust the balance between losses.

4 Experiment

The above model is trained by Monte Carlo-based EM algorithm proposed by [6,22]. In E-step, the hidden variable k_{t-1}^i is updated to k_t^i by Langevin Dynamics [23] in Snowmask Generator. Then, the M-step computes the total loss and achieves the optimization of the whole network parameter.

During the experiments, we use a Linux server equipped with GeForce RTX 3090 GPU and Pytorch. In training, we set the initial value of the learning rate as 1e-4, and it is halved every 20 epochs. Both synthetic and real snowy videos are clipped into patches of size 64 × 64. We train 100 epochs totally and the Adam algorithm is used to optimize the model parameters.

We use full-reference metrics (e.g., PSNR, SSIM) and non-reference metrics (e.g., NIQE, BRISQUE) to evaluate the results of the synthesized snowy video and real-world snowy video respectively. Furthermore, in order to verify the effectiveness of our method for subsequent vision tasks, we use the BoxMOT [24] to conduct further experiments of multiple object tracking.

4.1 Comparison with Existed Methods

In this section, we compare our method with advanced rain/snow removal methods based on deep learning in the past three years, including video deraining methods ESTINet [7], S2VD [6], video restoration method BasicVSR++ [25], as well as single image desnowing methods SnowFormer [26], HDCWNet [10]. Because of the poor performance of directly applying the deraining model to snowy videos, we retrain ESTINet, S2VD and BasicVSR++ with our dataset before testing.

Table 1. Comparison with state-of-the-art methods.

Methods	SRSVD				RVSD [8]	
	Synthetic Dataset		Real Dataset		Synthetic Dataset	
	PSNR↑	SSIM↑	NIQE↓	BRISQUE↓	PSNR↑	SSIM↑
ESTINet [7]	30.14	0.8981	3.8828	35.54	23.56	0.8614
S2VD [6]	34.74	0.9450	3.4196	37.71	22.95	0.8590
BasicVSR++ [25]	35.59	0.9426	**3.0407**	30.66	22.64	0.8618
SnowFormer [26]	20.20	0.7760	3.7469	33.85	24.01	**0.8939**
HDCWNet [10]	19.12	0.6963	3.6093	38.39	22.63	0.8592
ours	**36.57**	**0.9577**	3.3345	**30.19**	**24.89**	0.8756

We performed snow removal experiments on RVSD [8] and our dataset SRSVD. The quantitative results of the tests are illustrated in Table 1, and the best results are shown bold and the second best are underlined. For SRSVD, our method improves at least 0.98 dB PSNR and 0.0127 SSIM compared to other methods. Overall, the PSNR and SSIM metrics of methods based on video are generally better, which proves the importance of temporal correlation. On the real dataset, our method ranks first in BRISQUE and second in NIQE. For RVSD, our method also achieves good results.

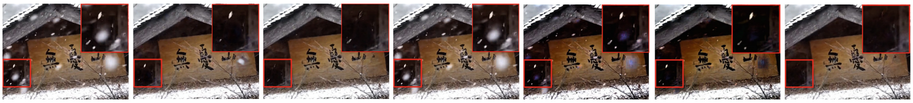


Fig. 5. Visual comparison of different methods on real-world snowy video of SRSVD.

The test results on the real dataset of SRSVD are shown in Fig. 5. As indicated in the red-boxed region, our method is the best to remove snow in this area. Through comparison, (d) has the worst performance, which may be because (d) only considers the synthetic snowy videos, resulting in poor generalization to desnow in real-world scenarios. In addition, methods(e)(f) based on single image are struggle to completely eliminate the snowflake and preserve the details due to the lack of interframe information. In contrast, The methods based on video(b)(c) and our method(g) generally outperform (e)(f). It proves that temporal correlation is significant to remove snow and supplement the background details occluded by snowflakes based on information from adjacent frames.

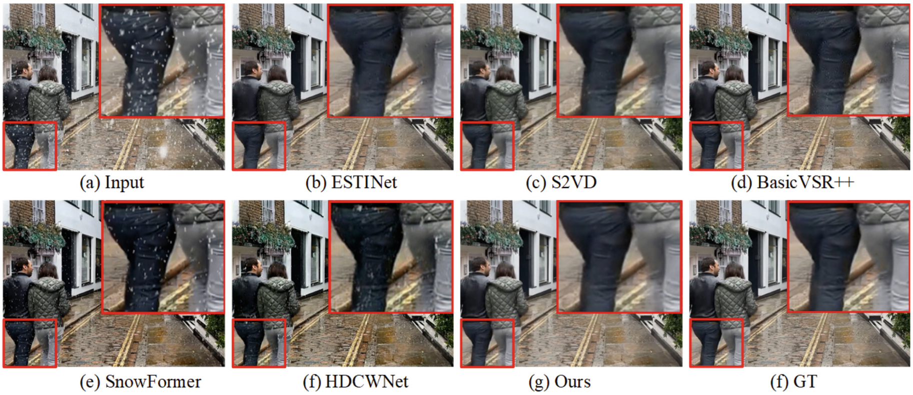

(a) Input
(b) ESTINet
(c) S2VD
(d) BasicVSR++
(e) SnowFormer
(f) HDCWNet
(g) Ours
(f) GT

Fig. 6. Visual comparison of different methods on synthetic snowy video of SRSVD.

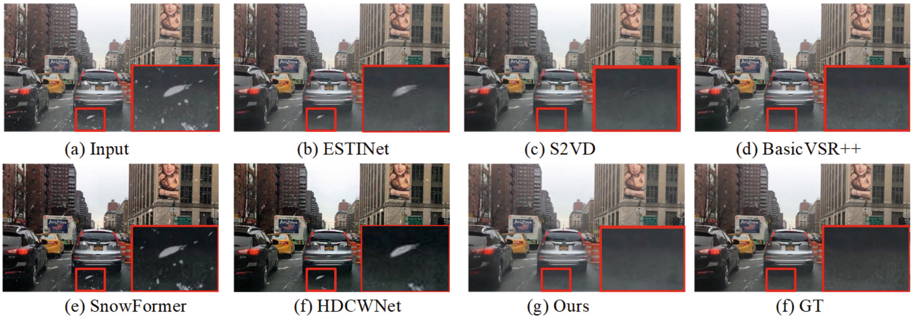

(a) Input
(b) ESTINet
(c) S2VD
(d) BasicVSR++
(e) SnowFormer
(f) HDCWNet
(g) Ours
(f) GT

Fig. 7. Visual comparison of different methods on synthetic snowy video of RVSD.

The results on the synthetic dataset of SRSVD are shown in Fig. 6. In the red-boxed region of (b)(c)(d), there are still a few small snow spots left. In particular, methods(e)(f) have trouble removing snow from pedestrians because of a lack of interframe information. It is obvious that our method outperforms other state-of-the-art methods on video desnowing.

The results on the synthetic dataset of RVSD are shown in Fig. 7. It shows that our method and (c)(d) have good snow removal performance, while (b)(e)(f) obviously have residual snow.

In addition, we conduct object tracking experiments on real dataset. Through the video/Image labeling and annotation tool named DarkLabel(ver2.4), the real snowy videos are labeled in the format of MOT17 [27]. Then the initial videos and the desnowing videos are input into BoxMOT [24] for testing. BoxMOT provides a great variety of tracking methods that meet different hardware limitations.

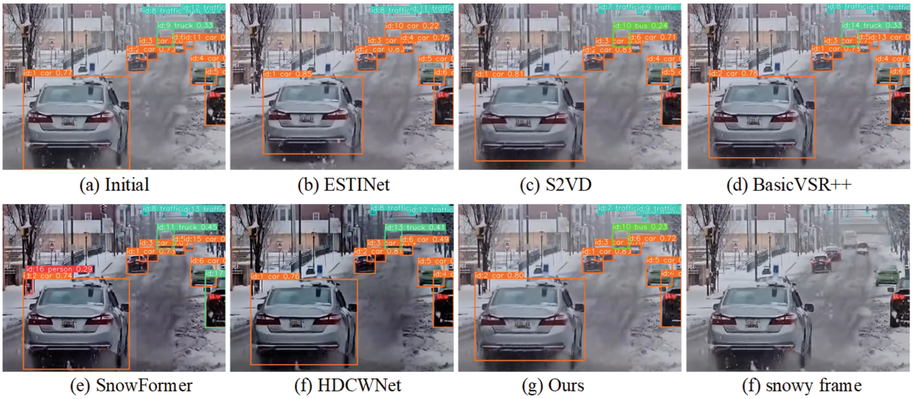


Fig. 8. Object tracking results on snowy and restored real-world video.

Table 2. Quantitative performance of object tracking on snowy and restored real-world video.

Input	HOTA↑	MOTA↑	IDF1↑
snowy video	36.01	29.32	38.17
ESTINet [7]	42.96	36.43	42.75
S2VD [6]	44.31	**38.65**	44.61
BasicVSR++ [25]	46.23	38.53	41.93
SnowFormer [26]	40.12	32.22	40.59
HDCWNet [10]	37.11	33.05	38.80
Ours	**47.34**	38.19	**44.89**

The experimental results on the real dataset are shown in Fig. 8. It can be seen that all methods except (c)(g) have residual snowflakes and misclassify bus as truck or car. Although our method misses one of the four traffic lights, it is still better than others methods. In Table 2, the MOTA(Multiple Object Tracking

Accuracy) [28], HOTA(Multiple Object Tracking Precision) [29] and IDF1(ID F1 score) [30] are used to evaluate the performance of object tracking. All three metrics are low for the original video due to snow occlusion. After desnowing, the tracking performance is improved. In particular, our method achieves the best scores on HOTA and IDF1.

4.2 Ablation Study

We take an ablation experiment to measure the effectiveness of the MSRB-P3D, the D-LKA 3Dblock and $\mathcal{L}_{interframe}$. To do so, we build a basic model $M1$ by removing $\mathcal{L}_{interframe}$ and D-LKA 3Dblock and replacing MSRB-P3D with residual blocks [31]. Then we add MSRB-P3D into M1 to construct M2 and add D-LKA 3Dblock into M2 to construct M3. Finally, $\mathcal{L}_{interframe}$ is added into M3 to construct our network.

As shown in Table 3, it reports that combing three components together has the best performance of video desnowing. Compared to the baseline network M1, our network improves the PSNR score from 34.68 to 36.57, the SSIM score from 0.9444 to 0.9577, the NIQE score from 3.7408 to 3.3345 and the BRISQUE score from 30.83 to 30.19. From the experiment of the synthetic dataset, it can be seen that MSRB-P3D, D-LKA 3Dblock and $\mathcal{L}_{interframe}$ improve the desnowing indicator score to some extent.

Table 3. Ablation study of different architectures in our work.

Index	MSRB-P3D	D-LKA 3Dblock	$\mathcal{L}_{interframe}$	SRSVD			
				Synthetic Dataset		Real Dataset	
				PSNR↑	SSIM↑	NIQE↓	BRISQUE↓
M1				34.68	0.9444	3.7408	30.83
M2	✓			35.28	0.9473	3.6740	31.00
M3	✓	✓		35.55	0.9506	3.6401	30.62
ours	✓	✓	✓	**36.57**	**0.9577**	**3.3345**	**30.19**

5 Conclusion

To solve the problems of incomplete snowflake removal and artifacts on moving objects in snowy video restoration, a deformable multi-scale video desnowing network is proposed in this paper. Experiments prove that our method not only outperforms other state-of-the-art methods on both synthetic and real snowy videos, but also effectively improves performance on subsequent vision task. In future work, we will further investigate video desnowing to improve the performance of subsequent vision task, such as building an end-to-end network.

References

1. Bossu, J., Hautiere, N., Tarel, J.P.: Rain or snow detection in image sequences through use of a histogram of orientation of streaks. Int. J. Comput. Vision **93**, 348–367 (2011)
2. Kim, J.H., Sim, J.Y., Kim, C.S.: Video deraining and desnowing using temporal correlation and low-rank matrix completion. IEEE Trans. Image Process. **24**(9), 2658–2670 (2015)
3. Li, M., et al.: Video rain streak removal by multiscale convolutional sparse coding. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 6644–6653 (2018)
4. Li, M., Cao, X., Zhao, Q., Zhang, L., Meng, D.: Online rain/snow removal from surveillance videos. IEEE Trans. Image Process. **30**, 2029–2044 (2021)
5. Ren, W., Tian, J., Han, Z., Chan, A., Tang, Y.: Video desnowing and deraining based on matrix decomposition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4210–4219 (2017)
6. Yue, Z., Xie, J., Zhao, Q., Meng, D.: Semi-supervised video deraining with dynamical rain generator. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 642–652 (2021)
7. Zhang, K., Li, D., Luo, W., Ren, W., Liu, W.: Enhanced spatio-temporal interaction learning for video deraining: faster and better. IEEE Trans. Pattern Anal. Mach. Intell. **45**(1), 1287–1293 (2022)
8. Chen, H., et al.: Snow removal in video: a new dataset and a novel method. In: 2023 IEEE/CVF International Conference on Computer Vision (ICCV), pp. 13165–13176. IEEE (2023)
9. Liu, Y.F., Jaw, D.W., Huang, S.C., Hwang, J.N.: DesnowNet: context-aware deep network for snow removal. IEEE Trans. Image Process. **27**(6), 3064–3073 (2018)
10. Chen, W.T., et al.: All snow removed: single image desnowing algorithm using hierarchical dual-tree complex wavelet representation and contradict channel loss. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 4196–4205 (2021)
11. Zhang, K., Li, R., Yu, Y., Luo, W., Li, C.: Deep dense multi-scale network for snow removal using semantic and depth priors. IEEE Trans. Image Process. **30**, 7419–7431 (2021)
12. Chen, X., et al.: Unpaired deep image deraining using dual contrastive learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 2017–2026 (2022)
13. Chen, W.-T., Fang, H.-Y., Ding, J.-J., Tsai, C.-C., Kuo, S.-Y.: JSTASR: joint size and transparency-aware snow removal algorithm based on modified partial convolution and veiling effect removal. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J.-M. (eds.) ECCV 2020. LNCS, vol. 12366, pp. 754–770. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-58589-1_45
14. Yang, W., Tan, R.T., Wang, S., Liu, J.: Self-learning video rain streak removal: when cyclic consistency meets temporal correspondence. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 1720–1729 (2020)
15. Wang, S., Zhu, L., Fu, H., Qin, J., Schönlieb, C.B., Feng, W., Wang, S.: Rethinking video rain streak removal: a new synthesis model and a deraining network with video rain prior. In: European Conference on Computer Vision, pp. 565–582. Springer (2022)

16. Yang, Y., Aviles-Rivero, A.I., Fu, H., Liu, Y., Wang, W., Zhu, L.: Video adverse-weather-component suppression network via weather messenger and adversarial backpropagation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 13200–13210 (2023)
17. Xue, T., Zhou, G., He, R., Wang, Z., Chen, J., Jia, Z.: RVDNet: a two-stage network for real-world video desnowing with domain adaptation. In: ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 3305–3309. IEEE (2024)
18. Christiansen, M.: Adobe after effects CC visual effects and compositing studio techniques. Adobe Press (2013)
19. Azad, R., et al.: Beyond self-attention: deformable large kernel attention for medical image segmentation. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pp. 1287–1297 (2024)
20. Tong, Z., Song, Y., Wang, J., Wang, L.: VideoMAE: masked autoencoders are data-efficient learners for self-supervised video pre-training. Adv. Neural. Inf. Process. Syst. **35**, 10078–10093 (2022)
21. Ying, X., Wang, L., Wang, Y., Sheng, W., An, W., Guo, Y.: Deformable 3d convolution for video super-resolution. IEEE Signal Process. Lett. **27**, 1500–1504 (2020)
22. Dempster, A.P., Laird, N.M., Rubin, D.B.: Maximum likelihood from incomplete data via the EM algorithm. J. Roy. Stat. Soc.: Ser. B (Methodol.) **39**(1), 1–22 (1977)
23. Uhlenbeck, G.E., Ornstein, L.S.: On the theory of the Brownian motion. Phys. Rev. **36**(5), 823 (1930)
24. Jonathon Luiten, A.H.: Trackeval (2020). https://github.com/JonathonLuiten/TrackEval
25. Chan, K.C., Zhou, S., Xu, X., Loy, C.C.: BasicVSR++: improving video super-resolution with enhanced propagation and alignment. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 5972–5981 (2022)
26. Chen, S., Ye, T., Liu, Y., Chen, E., Shi, J., Zhou, J.: SnowFormer: Scale-aware transformer via context interaction for single image desnowing. arXiv preprint arXiv:2208.09703 (2022)
27. Milan, A., Leal-Taixé, L., Reid, I., Roth, S., Schindler, K.: MOT16: A benchmark for multi-object tracking. arXiv:1603.00831 [cs] (2016). http://arxiv.org/abs/1603.00831, arXiv: 1603.00831
28. Bernardin, K., Stiefelhagen, R.: Evaluating multiple object tracking performance: the clear mot metrics. EURASIP J. Image Video Process. **2008**, 1–10 (2008)
29. Luiten, J., Osep, A., Dendorfer, P., Torr, P., Geiger, A., Leal-Taixé, L., Leibe, B.: HOTA: a higher order metric for evaluating multi-object tracking. In: International Journal of Computer Vision, pp. 1–31 (2020)
30. Ristani, E., Solera, F., Zou, R., Cucchiara, R., Tomasi, C.: Performance measures and a data set for multi-target, multi-camera tracking. In: Hua, G., Jégou, H. (eds.) ECCV 2016. LNCS, vol. 9914, pp. 17–35. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-48881-3_2
31. He, K., Zhang, X., Ren, S., Sun, J.: Identity mappings in deep residual networks. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9908, pp. 630–645. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46493-0_38

Fast Orthogonal Matching Pursuit Through Successive Regression

Huiyuan Yu, Jia He, and Maggie Cheng(✉)

Illinois Institute of Technology, Chicago, IL 60616, USA
{hyu47,jhe58}@hawk.iit.edu, maggie.cheng@iit.edu

Abstract. Orthogonal Matching Pursuit (OMP) has been a powerful method in sparse signal recovery and approximation. However, OMP suffers computational issues when the signal has a large number of non-zeros. This paper advances OMP and its extension called generalized OMP (gOMP) by offering fast algorithms for the orthogonal projection of the input signal at each iteration. The proposed modifications directly reduce the computational complexity of OMP and gOMP. Experiment results verified the improvement in computation time. This paper also provides sufficient conditions for exact signal recovery. For general signals with additive noise, the approximation error is at the same order as OMP (gOMP), but is obtained within much less time.

Keywords: Greedy Algorithm · Compressive Sensing · Sparse Signal Recovery · Approximation · Orthogonal Matching Pursuit

1 Introduction

Let $\boldsymbol{x}$ be a d-dimensional real signal. Suppose there is a real measurement matrix $\Phi \in \mathbb{R}^{N\times d}$, through which we can obtain an N-dimensional measurement $\boldsymbol{y} = \Phi\boldsymbol{x}$. Usually $N < d$, which presents an underdetermined system. How to reconstruct the original signal $\boldsymbol{x}$ from an underdetermined system? If $\boldsymbol{x}$ is sparse, then by exploiting sparsity, we may be able to find a unique solution. $\boldsymbol{x}$ is called a k-sparse signal if $\boldsymbol{x}$ has at most k non-zero components.

The measurement matrix Φ is also called a dictionary, and each column $\boldsymbol{\varphi}$ of the dictionary called an atom. Let $\mathcal{J} = \{1, \ldots, d\}$ represent the index set of all atoms in the dictionary. If the dictionary is overcomplete, there are many representations of $\boldsymbol{y} = \sum_{\gamma\in\mathcal{J}} a_\gamma \boldsymbol{\varphi}_\gamma$. Intuitively, we would like to find the sparsest solution: $\min_{\boldsymbol{x}} \|\boldsymbol{x}\|_0$ subject to $\boldsymbol{y} = \Phi\boldsymbol{x}$, but it is an NP-hard problem. Different optimization principles lead to different sparse representations of $\boldsymbol{y}$, for example, basis pursuit (BP) [5,6,12] and the method of frames (MOF) [8] among many others [13,28]:

- Find a representation of the input signal whose coefficients have the minimal ℓ_1 norm.

$$\min_{\boldsymbol{x}} \|\boldsymbol{x}\|_1 \quad \text{subject to} \quad \boldsymbol{y} = \Phi\boldsymbol{x} \tag{BP}$$

A. Antonacopoulos et al. (Eds.): ICPR 2024, LNCS 15332, pp. 302–315, 2025.
https://doi.org/10.1007/978-3-031-78125-4_21

- Find a representation of the input signal whose coefficients have the minimal ℓ_2 norm.

$$\min_{\boldsymbol{x}} \|\boldsymbol{x}\|_2 \quad \text{subject to} \quad \boldsymbol{y} = \boldsymbol{\Phi x} \tag{MOF}$$

BP and MOF both provide convex relaxation to the ℓ_0 norm minimization problem, however, neither of them provides the sparsest solution, except those satisfying the sparsity condition specified in [11].

Matching Pursuit (MP) [18] uses an iterative procedure that directly addresses the sparsity issue. Orthogonal Matching Pursuit (OMP) [21,25,26] inherits the greedy approach from MP that selects an atom with the maximal correlation with the residual at present, but improves over the standard MP by adding least square minimization at each iteration. Let Γ be the index set of atoms found so far, the least square estimation is used for computing the orthogonal projection of the input signal $\boldsymbol{y}$ onto the subspace spanned by the atoms indexed by Γ:

$$\min_{\boldsymbol{x}_\Gamma} \|\boldsymbol{y} - \boldsymbol{\Phi}_\Gamma \boldsymbol{x}_\Gamma\|_2^2 \quad \text{with } |\Gamma| \leq k \tag{OMP}$$

OMP has been shown to have better results than MP. Many variations of OMP have been developed [7,10,17,19,20,27]. Under certain conditions OMP provides recovery guarantee [2–4,9,25,26]. The excellent performance of OMP results from the orthogonal projection of $\boldsymbol{y}$ onto the subspace spanned by the atoms selected so far. The least square solution is obtained by $\boldsymbol{x}_\Gamma = \boldsymbol{\Phi}_\Gamma^+ \boldsymbol{y}$. As Γ increases, solving the least square problem significantly increases the computational load. In this paper, we propose a fundamental improvement over classical OMP to avoid the high complexity of computing pseudo inverse over an increasing-sized matrix, which can be generalized to other OMP-based algorithms:

- When solving the least square problem at each iteration, instead of computing $\boldsymbol{\Phi}_\Gamma^+ \boldsymbol{y}$ over the entire support Γ, it uses successive regression over a single atom. It makes the same greedy choice as OMP does at each iteration, but is much faster due to reduced computation load. The proposed algorithm is called OMP-SR.
- The blocked version of OMP is called Generalized Orthogonal Matching Pursuit [27], which extends the greedy choice to multiple atoms at each iteration but still preserve the convergence property of OMP. We propose a blocked version of OMP-SR, called Blocked Successive Regression (BSR). BSR is an improvement over gOMP, analagous to OMP-SR being an improvement over OMP.

In general, the measurement $\boldsymbol{y}$ is often with noise. A general signal may be represented as the linear combination of atoms from the dictionary with additive noise,

$$\boldsymbol{y} = \boldsymbol{\Phi x} + \boldsymbol{\varepsilon}.$$

We are interested in the best approximation of $\boldsymbol{y}$ using a linear combination of atoms. *The best approximation* is the one with the smallest approximation error measured by the ℓ_2 norm of the residual, and hence, the optimization principle is,

$$\min_{\boldsymbol{x}} \|\boldsymbol{y} - \Phi\boldsymbol{x}\|_2 \quad \text{subject to } \|\boldsymbol{x}\|_0 \leq k \qquad \text{(Sparse Approximation)}$$

OMP is a fundational approach for signal reconstruction, therefore, any direct improvement over OMP can benefit many applications that use various implementation of OMP. The proposed method is also different from previous efforts that use matrix factorization based solutions (e.g., [22,29]) and matrix inversion bypass (MIB) technique ([14,15]). In [24], a simulation-based comparison have been provided over various implementation of OMP. In this paper we not only provide simulation based comparison, but also analytical complexity analysis.

The rest of the paper is organzied as follows: In Sect. 2, we present our algorithms for exact recover; in Sect. 3, we show the main theoretical results for the BSR algorithm[1]; in Sect. 4, we show the performance of our algorithms in real datasets compared to the baseline methods OMP and gOMP.

1.1 Notation

- $\Phi^\top$: transpose of matrix Φ
- Φ^+: pseudo inverse of matrix Φ
- $(\Phi^\top\Phi)^{-1}$: inverse of matrix $(\Phi^\top\Phi)$
- $\|A\|_{p\to q} = \sup\limits_{\boldsymbol{x}\neq 0} \frac{\|A\boldsymbol{x}\|_q}{\|\boldsymbol{x}\|_p}$: operator norm of matrix A.
- $\|A\|_{p\to p}$ is abbreviated to $\|A\|_p$.

2 Recovery Algorithms by Successive Regression

2.1 Orthogonal Matching Pursuit Through Successive Regression (OMP-SR)

OMP-SR is a fast implementation of OMP. When solving the least square problem at each iteration of OMP, it avoids the expensive computation for the pseudo inverse of $\Phi_{J^{t-1}}$; instead, it only projects onto the atom selected in the current iteration via univariate regression, and then updates the coefficients of atoms selected in previous iterations through a backtracking procedure: $b_l = \beta_l - \sum\limits_{k=l+1}^{t} b_k \gamma_{l,k}$(see Algorithm 1), where $\beta_t = \frac{\langle \boldsymbol{z}_t, \boldsymbol{y}\rangle}{\langle \boldsymbol{z}_t, \boldsymbol{z}_t\rangle}$ is the coefficient newly obtained in the current iteration, b_l is the updated coefficient for the atoms selected in previous iterations. Note that the inner product $\langle \boldsymbol{z}_l, \boldsymbol{z}_l\rangle$ does not need to be recomputed. It only needs to be computed once, that is when we compute β_l in the l-th iteration.

[1] The proofs for theorems and lemmas can be found in arXiv:2404.00146.

Algorithm 1. OMP-SR

```
Initialization: $\boldsymbol{a}_0 = \boldsymbol{z}_0 = \mathbf{1}$, $\boldsymbol{r}^0 = \boldsymbol{y}$, $J^0 = \phi$
for $t = 1$ to $\kappa$ do
    Choose $j^t = \arg\max_{j \in \mathcal{J} \setminus J^{t-1}} |\langle \boldsymbol{\varphi}_j, \boldsymbol{r}^{t-1} \rangle|$        ▷ $\boldsymbol{\varphi}_j$ is the $j$-th column of $\Phi$
    Let $\boldsymbol{a}_t$ be the $j^t$-th column of matrix $\Phi$.
    Regress $\boldsymbol{a}_t$ on $\boldsymbol{z}_l$ and get coefficients
        $$\gamma_{l,t} = \frac{\langle \boldsymbol{z}_l, \boldsymbol{a}_t \rangle}{\langle \boldsymbol{z}_l, \boldsymbol{z}_l \rangle}, \quad \text{for } l = 0, \ldots, t-1$$
    Compute $\boldsymbol{z}_t = \boldsymbol{a}_t - \sum_{l=0}^{t-1} \gamma_{l,t} \boldsymbol{z}_l$
    Regress $\boldsymbol{y}$ on $\boldsymbol{z}_t$ to get $\beta_t = \frac{\langle \boldsymbol{z}_t, \boldsymbol{y} \rangle}{\langle \boldsymbol{z}_t, \boldsymbol{z}_t \rangle}$
    Let $b_t = \beta_t$
    if $t > 1$ then
        for $l = t - 1$ to $1$ do
            $b_l = \beta_l - \sum_{k=l+1}^{t} b_k \gamma_{l,k}$
        end for
    end if
    Update index set $J^t = J^{t-1} \bigcup \{j^t\}$
    Update residual $\boldsymbol{r}^t = \boldsymbol{y} - \sum_{l=1}^{t} b_l \boldsymbol{\varphi}_{j^l}$
end for
Let $\boldsymbol{x}_{j^t} = b_t$ for $t = 1, \ldots, \kappa$, and let $\boldsymbol{x}_j = 0$ for $j \notin J^\kappa$
Return $\boldsymbol{x}$
```

OMP-SR selects the same atom and generates the same residual as OMP does at each iteration, and therefore returns the same result as OMP. OMP-SR starts to show performance gain over OMP when the number of non-zeros in $\boldsymbol{x}$ increases due to not having to compute the pseudo inverse of a growing matrix.

2.2 Complexity Comparison with QR-Based OMP

In practice, OMP implementation based on incremental QR decomposition may be used for improved efficiency (e.g., [1,16,22,23,29]). In each iteration, Q_t and R_t matrices are updated as in the algorithm. To obtain the updated solution for the least square problem, one needs to compute $\boldsymbol{h} = Q_t^\top \boldsymbol{y}$, and then use back-substitution to solve $R_t \boldsymbol{x} = \boldsymbol{h}$. However, despite the cost saving over standard OMP, the operation cost of OMP based on QR decomposition is still higher than that of the proposed OMP-SR. Table 1 and Table 2 show the floating-point operations of them for each iteration of the OMP algorithm.

A term-by-term comparison shows OMP-SR uses fewer flops than QR-based OMP. The cost analysis is for $\Phi \in \mathbb{R}^{N \times d}$. For sparse signals with k non-zeros, as long as $t(2N-1) < (d-t)(4N-1)$, OMP-SR outperforms QR-based OMP by a margin of at least $N+2$ per iteration. Typically in OMP, $t \leq k \ll d$ for

Table 1. Operation cost for the t-th iteration of OMP using QR update

Operation	Flops
Update Q_t, R_t	$(d-t)(4N-1)+3N+1$
Update $\boldsymbol{h} = Q_t^\top \boldsymbol{y}$	$2N$
Solve $R_t\boldsymbol{x} = \boldsymbol{h}$	t^2
Total cost	$(d-t)(4N-1)+5N+1+t^2$

Table 2. Operation cost for the t-th iteration of the proposed OMP-SR

Operation	Flops
Compute $\gamma_{l,t}$	$2Nt$
Compute β_t	$4N-1$
Update coefficients b_l	t^2-t
Total cost	$t(2N-1)+4N-1+t^2$

sparse recovery problems, therefore, the condition $t(2N-1) < (d-t)(4N-1)$ is easily satisfied.

2.3 Blocked Successive Regression (BSR)

BSR builds on the idea of successive regression in OMP-SR but selects a block of atoms at each iteration. The block size c is a hyper parameter, usually decided by a grid search. The algorithm is still greedy in nature: in each iteration it selects the atoms that have the largest correlations with the residual measured by the ℓ_2 norm. Each iteration of BSR performs an orthogonal projection of $\boldsymbol{y}$ over c newly selected atoms, instead of over $|\mathit{\Gamma}|$ atoms, which could be costly as $|\mathit{\Gamma}|$ increases with iterations. Subsequently the coefficients for atoms selected in previous iterations are updated through $b_i = \beta_i - \sum\limits_{k=l+1}^{t} \sum\limits_{j\in\mathit{\Gamma}_k} b_j\gamma_{i,j}$ (see Algorithm 2).

The BSR algorithm halts if the residual becomes too small or it has exhausted κ iterations, which amounts to two of the three halting rules listed in [19] for matching pursuit type of algorithms.

The columns selected by BSR shall be the same as the columns selected by gOMP [27] in each iteration. However, the two algorithms differ in the way they solve the least square problem.

3 Conditions for Exact Recovery

3.1 Background

Assume there are k non-zero entries in a d-dimensional signal $\boldsymbol{x}$, and $k \ll d$. Let $\mathit{\Lambda}_{\text{opt}} = \{i_1, \ldots, i_k\}$ be the set of indices for the non-zero entries of $\boldsymbol{x}$. Without

Algorithm 2. BSR

```
Initialization: $\boldsymbol{r}^0 = \boldsymbol{y}$, $\Gamma = \phi$, $\boldsymbol{z}_0 = \boldsymbol{1}$
for $t = 1$ to $\kappa$ do
    $\Gamma_t = \arg \max_{\substack{\Omega:|\Omega|=c \\ \Omega \subset \mathcal{J} \setminus \Gamma}} \left\| \Phi_\Omega^\top \boldsymbol{r}^{t-1} \right\|_2$
    for each $j \in \Gamma_t$ do
        Let $\boldsymbol{a}_j$ be the $j$-th column of matrix $\Phi$.
        Regress $\boldsymbol{a}_j$ on $\boldsymbol{z}_0$ to get coefficient $\gamma_{0,j} = \frac{\langle \boldsymbol{z}_0, \boldsymbol{a}_j \rangle}{\langle \boldsymbol{z}_0, \boldsymbol{z}_0 \rangle}$
        Compute $\boldsymbol{z}_j = \boldsymbol{a}_j - \gamma_{0,j} \boldsymbol{z}_0$
        if $t > 1$ then
            Regress $\boldsymbol{a}_j$ on $Z_{\Gamma_l}$ to get coefficients
                $\boldsymbol{\gamma}_{\Gamma_l,j} = Z_{\Gamma_l}^+ \boldsymbol{a}_j$, for $l = 1, \ldots, t-1$
            Compute $\boldsymbol{z}_j = \boldsymbol{z}_j - \sum_{l=1}^{t-1} \sum_{i \in \Gamma_l} \gamma_{i,j} \boldsymbol{z}_i$
        end if
    end for
    Regress $\boldsymbol{y}$ on $Z_{\Gamma_t}$ to get coefficients $\boldsymbol{\beta}_{\Gamma_t} = Z_{\Gamma_t}^+ \boldsymbol{y}$
    Let $\boldsymbol{b}_{\Gamma_t} = \boldsymbol{\beta}_{\Gamma_t}$
    if $t > 1$ then
        for $l = t-1$ to $1$ do
            for $i \in \Gamma_l$ do
                $b_i = \beta_i - \sum_{k=l+1}^{t} \sum_{j \in \Gamma_k} b_j \gamma_{i,j}$
            end for
        end for
    end if
    Update index set $\Gamma = \Gamma \bigcup \Gamma_t$
    Update residual $\boldsymbol{r}^t = \boldsymbol{y} - \Phi_\Gamma \boldsymbol{b}_\Gamma$
    Break if $\|\boldsymbol{r}^t\|_2 \leq \delta$
end for
Let $\boldsymbol{x}_\Gamma = \boldsymbol{b}_\Gamma$, and let $\boldsymbol{x}_{\mathcal{J} \setminus \Gamma} = 0$
Return $\boldsymbol{x}$
```

loss of generality, we can partition the measurement matrix as $\Phi = [\Phi_{\text{opt}} | \Psi]$ so that Φ_{opt} has k columns, $\Phi_{\text{opt}} = [\boldsymbol{\varphi}_{i_1}, \ldots, \boldsymbol{\varphi}_{i_k}]$, and Ψ has the remaining $d-k$ columns.

In the absence of noise, the measured signal $\boldsymbol{y}$ has a sparse representation: $\boldsymbol{y} = \Phi \boldsymbol{x} = \sum_{j \in \Lambda_{\text{opt}}} a_j \boldsymbol{\varphi}_j$. Exact recovery aims to recover the coefficients a_j for all atoms indexed by Λ_{opt}, which are the non-zero entries in $\boldsymbol{x}$.

The Exact Recovery Condition of OMP-SR. Algorithm OMP-SR essentially is an OMP algorithm with fast implementation: it starts with the same initial residual $\boldsymbol{r}^0$ and selects the same atom in the next iteration, so the residual $\boldsymbol{r}^t$ after the t-th iteration is the same. Since $\boldsymbol{r}^t$ is used as input to the next iteration when choosing a column, the next iteration will result in the same residual $\boldsymbol{r}^{t+1}$. By induction, after k iterations the algorithm returns the same result as OMP does. The exact recovery condition for OMP-SR is the same as for OMP.

3.2 The Exact Recovery Conditions of BSR

BSR is essentially a greedy algorithm, which makes a greedy choice at each iteration, except that BSR selects a block of columns at each iteration with a fixed block size c ($c \geq 1$). If $c = 1$, BSR reduces to OMP-SR. We have learned that under the condition of $\rho(\boldsymbol{r}) < 1$, OMP and OMP-SR can find one optimal column in each iteration. Then for BSR, under what condition will each iteration of BSR only select the optimal columns from Φ_{opt} except the last iteration? This is the best case, in which BSR can locate all optimal columns within $\lceil k/c \rceil$ iterations. We call the condition for the best case as *the strong exact recovery condition for BSR.*

A Strong Exact Recovery Condition for BSR. Recall that $\Phi = [\Phi_{\text{opt}}|\Psi]$ so that Φ_{opt} has the k optimal columns, and Ψ has the remaining $d-k$ columns. Let $\boldsymbol{r}$ denote the residual at the current iteration before the greedy choice is made.

For a fixed block size c, the greedy choice ratio is defined as follows:

$$\rho_c(\boldsymbol{r}) \overset{\text{def}}{=\!=} \frac{\max\limits_{\Omega_1} \left\|\Phi_{\Omega_1}^{\top}\boldsymbol{r}\right\|_2}{\max\limits_{\Omega_2} \left\|\Phi_{\Omega_2}^{\top}\boldsymbol{r}\right\|_2}, \tag{1}$$

such that $|\Omega_1| = |\Omega_2| = c$ and $|\Omega_2 \cap \Lambda_{\text{opt}}| > |\Omega_1 \cap \Lambda_{\text{opt}}|$, i.e., Ω_2 has at least one more optimal column than Ω_1. Given a k-sparse signal, BSR can recover the signal within $\lceil k/c \rceil$ iterations if the following condition holds.

Theorem 1 (The strong exact recovery condition for BSR). *A sufficient condition for BSR to recover a k-sparse signal within $\lceil k/c \rceil$ iterations is that*

$$\rho_c(\boldsymbol{r}) < 1 \tag{2}$$

holds for all iterations.

A Weak Exact Recovery Condition for BSR. What is the condition for $\rho_c(\boldsymbol{r}) < 1$ to hold in Theorem 1? In the absence of a straightforward answer, we first discuss the condition for BSR to recover a k-sparse signal within k iterations, then revisit the condition (2).

We call the condition for BSR to recover a k-sparse signal within k iterations *the weak exact recovery condition for BSR.* For the weak condition, we use the following greedy choice ratio: $\rho(\boldsymbol{r}) = \frac{\left\|\Psi^{\top}\boldsymbol{r}\right\|_\infty}{\left\|\Phi_{\text{opt}}^{\top}\boldsymbol{r}\right\|_\infty}$.

Theorem 2 (The Weak Exact Recovery Condition for BSR). *A sufficient condition for BSR to recover a k-sparse signal within k iterations is that*

$$\rho(\boldsymbol{r}) < 1 \tag{3}$$

holds for all iterations.

Although intuitive, the condition in (3) expressed in terms of the greedy choice ratio cannot be checked before we know the residuals in all iterations. We need to establish a sufficient condition for the exact recovery by BSR in terms of the property of the dictionary Φ.

BSR will select at least one optimal column at each iteration but can also possibly select some non-optimal columns. We can split the columns of Ψ into two parts: $\Psi = [\Psi_J | \Psi_{\overline{J}}]$, where Ψ_J are the non-optimal columns that have been selected by BSR algorithm so far, and $\Psi_{\overline{J}}$ include the remaining columns.

Let matrix X be the submatrix of Φ that includes all columns of Φ_{opt} and the columns in Ψ that have been selected by BSR at the previous iterations, i.e., $X = [\Phi_{\mathrm{opt}} | \Psi_J]$.

Let Π denote the index set for the columns in Φ_{opt} that have not been selected by the algorithm so far, so $|\Pi| \leq k$. Let $(\cdot)_\Pi$ denote the columns in the matrix indexed by Π, and $(\cdot)_{\Pi,:}$ denotes the rows of the matrix indexed by Π.

Lemma 1. *If* $\max\limits_{\boldsymbol{\psi}} \left\| (X^+)_{\Pi,:} \boldsymbol{\psi} \right\|_1 < 1$*, where vector* $\boldsymbol{\psi}$ *ranges over columns of* $\Psi_{\overline{J}}$*, then the residual* $\boldsymbol{r}$ *satisfies* $\rho(\boldsymbol{r}) < 1$*.*

Although condition $\max\limits_{\boldsymbol{\psi} \in \Psi_{\overline{J}}} \left\| (X^+)_{\Pi,:} \boldsymbol{\psi} \right\|_1 < 1$ is expressed in terms of the property of the dictionary, this condition still cannot be checked without executing the algorithm. In practice it is unlikely that the optimal columns are known *a priori*, so the submatrices $X, \Psi_{\overline{J}}$ cannot be located before the execution of the algorithm. More practical methods are needed to check the sufficient condition without the execution of the algorithm.

In [25], a fundamental property of the dictionary Φ, called *coherence* is defined as:

$$\mu \overset{\mathrm{def}}{=} \max_{j \neq k} |\langle \boldsymbol{\varphi}_j, \boldsymbol{\varphi}_k \rangle| \tag{4}$$

Coherence μ is the maximum absolute value of pairwise inner product between the columns of the dictionary.

For a positive integer m, the cumulative coherence function, $\mu_1(m)$ of the dictionary, is defined as

$$\mu_1(m) \overset{\mathrm{def}}{=} \max_{|\Lambda|=m} \max_{\boldsymbol{\psi}} \sum_{j \in \Lambda} |\langle \boldsymbol{\varphi}_j, \boldsymbol{\psi} \rangle| \tag{5}$$

where Λ is the set of indices for any m columns of Φ, and $\boldsymbol{\psi}$ ranges over the columns of Φ not indexed by Λ. $\mu_1(m)$ is the maximum cumulative coherence from any m columns of Φ.

Next, we use the cumulative coherence property of the dictionary to derive a sufficient condition.

Lemma 2. $\max_{\psi\in\Psi_{\bar{J}}} \left\|(X^{+})_{\Pi,:}\boldsymbol{\psi}\right\|_1 < 1$ *whenever* $\mu_1(l)+\mu_1(n)<1$ *holds, where* n *is the number of columns in* X*, and* $l=\min(|\Pi|,k-1)$.

Lemma 2 and Lemma 1 together lead to the following conclusion: the residual $\boldsymbol{r}$ satisfies $\rho(\boldsymbol{r})<1$ whenever

$$\mu_1(l)+\mu_1(n)<1. \tag{6}$$

Revisit Theorem 1: the Sufficient Condition for $\boldsymbol{\rho_c}(\mathbf{r})<\mathbf{1}$. It is easy to show that $\mu_1(l)+\mu_1(n)<1$ is also sufficient for $\rho_c(\boldsymbol{r})<1$ to hold in Theorem 1, which leads to the following theorem.

Theorem 3 (The strong exact recovery condition for BSR). *Suppose that* μ *is the coherence of the dictionary as defined in* (4)*. A sufficient condition for BSR to recover a* k*-sparse signal within* $\lceil k/c\rceil$ *iterations is that*

$$\mu(2k-1)<1. \tag{7}$$

4 Experiments

Data used in the experiments are posted at github.

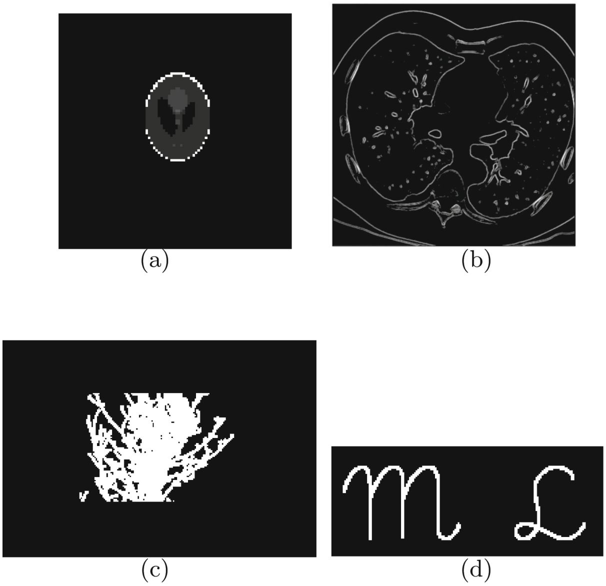
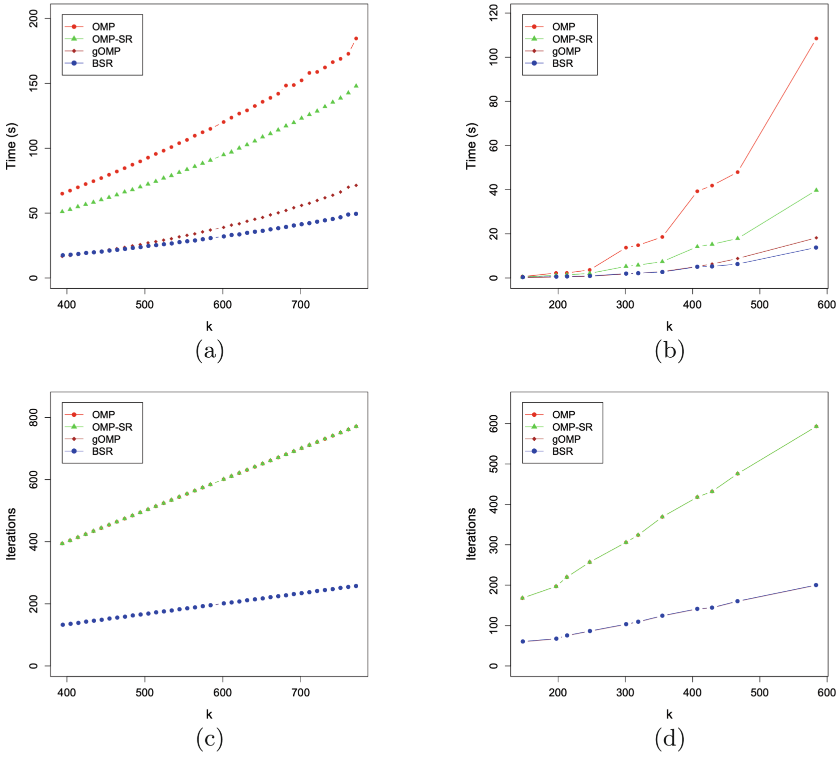

(a) (b) (c) (d)

Fig. 1. Images used for experiments, (a) phantom, (b) transaxial CT, (c) trees, (d) letters.

4.1 Sparse Signal Recovery

We first show sparse signal recovery performance when the signal has a sparse representation. The first experiment is on image data, where the non-zero elements constitute the content of an image, and exhibit continuity in the true signal $\boldsymbol{x}$. The images we used include the phantom, a CT scan, an image for trees, an image for letters (see Fig. 1), and MNIST dataset handwritten digits. The second experiment is on signals defined on graph structures, where the non-zero elements are distributed among the nodes of a graph. We used synthetic data defined on a binary tree, and data that are collected from IEEE 118-bus power system and IEEE 1354-bus power system, where the true signal $\boldsymbol{x}$ consists of the values of the state variables of a power system. Since the algorithms do not depend on the signal structure to find the non-zeros, they worked well with both types of data. Table 3 and Table 4 show the performance of the proposed OMP-SR and BSR, and we report the number of iterations, the recovered optimal atoms, normalized MSE (NMSE), and running time in seconds. Image data are reported in Table 3, and graph data are reported in Table 4.



(a) (b)

(c) (d)

Fig. 2. Running time (top) and iterations (bottom) used by the algorithms to recover k non-zeros in the signal. OMP and OMP-SR use the same number of iterations, and gOMP and BSR use the same number of iterations. The datasets used: (a),(c) phantom; (b),(d) MNIST dataset handwritten digit '7'.

If a k-sparse signal can be recovered by OMP within k iterations, it can certainly be recovered by BSR within k iterations. Those that cannot be recovered by OMP within k iterations are shown to take far less than k iterations and far less time by the BSR algorithm to fully recover. Since OMP-SR picks the same atoms as OMP does, we reported the result of OMP-SR in the same row as OMP and only reported its time (in blue text). Similarly, we report BSR in the same row as gOMP, and report its running time (in blue text). It is observed that OMP-SR is faster than OMP, and BSR is faster than gOMP. The block size c in BSR is a hyper parameter searched from $\{2, 3, 4, 8\}$.

The third experiment is to show the relation between k and running time. Image data for the phantom and the MNIST handwritten digit '7' were used. We created different versions from the original image to have different image sizes d and different k/d ratios. Figure 2 shows how the running time and iteration number increase as the number of non-zeros k increases. The number of iterations is reduced by several folds in the blocked version, which is shown in (c) and (d). BSR is faster than gOMP per iteration, however, due to the reduced number of iterations, the advantage of BSR over gOMP becomes less significant compared

Table 3. Image datasets. Reported NMSE and time in seconds. Running time of our methods is highlighted in blue. OMP-SR is faster than OMP, and BSR is faster than gOMP.

Data	k	method	ite	found	NMSE	time
MNIST(3) 392×784	126	OMP (OMP-SR)	126	110	0.0656	0.4077 (0.2977)
		OMP (OMP-SR)	142	126	$<1e{-}11$	0.5799 (0.3280)
		gOMP (BSR)	51	126	$<1e{-}11$	0.2167 (0.1298)
MNIST(5) 392×784	162	OMP (OMP-SR)	162	99	0.7058	0.5902 (0.4204)
		OMP (OMP-SR)	784	162	$<1e{-}11$	18.0441 (12.9938)
		gOMP (BSR)	85	162	$<1e{-}11$	0.5632 (0.3926)
MNIST(8) 392×784	174	OMP (OMP-SR)	174	92	0.8275	0.7703 (0.4934)
		OMP (OMP-SR)	546	174	$<1e{-}11$	11.0087 (6.5816)
		gOMP (BSR)	84	174	$<1e{-}11$	0.5570 (0.3855)
MNIST(9) 392×784	130	OMP (OMP-SR)	130	125	0.0727	0.4453 (0.3190)
		OMP (OMP-SR)	135	130	$<1e{-}11$	0.5620 (0.3917)
		gOMP (BSR)	35	130	$<1e{-}11$	0.1521 (0.1111)
Phantom 4512×9024	641	OMP (OMP-SR)	641	638	0.0278	89.0610 (70.8633)
		OMP (OMP-SR)	644	641	$<1e{-}11$	97.9924 (75.9530)
		gOMP (BSR)	81	641	$<1e{-}11$	13.0361 (9.1253)
Transaxial CT 4225×8450	1089	OMP (OMP-SR)	1089	1064	0.0675	282.6071 (250.1557)
		OMP (OMP-SR)	1115	1089	$<1e{-}11$	302.0011 (263.0615)
		gOMP (BSR)	57	1089	$<1e{-}11$	17.5773 (14.0619)
Trees 19200×38400	4670	OMP (OMP-SR)	4670	4652	0.00114	2391.0661 (702.4279)
		OMP (OMP-SR)	4688	4670	$<1e{-}11$	2571.1333 (754.8234)
		gOMP (BSR)	117	4670	$<1e{-}11$	23.0564 (20.0294)
Letters 5712×11424	851	OMP (OMP-SR)	851	851	$<1e{-}11$	191.7733 (129.9197)
		gOMP (BSR)	107	851	$<1e{-}11$	20.0811 (16.4433)

to the advantage of OMP-SR over OMP, as the iteration number is reduced significantly.

4.2 Sparse Approximation

The fourth experiment is for the sparse approximation of general signals with noises. We add noise $\boldsymbol{\varepsilon}$ to the Phantom image, and report approximation errors when the measurements are subject to increasing levels of noise. Table 5 shows that at each noise level, BSR found the k non-zeros with fewer iterations than OMP and significantly less running time.

Table 4. Synthetic data for signals defined on graph structures. Reported NMSE and time in seconds. Running time of our methods is highlighted in blue. OMP-SR is faster than OMP, and BSR is faster than gOMP.

Data	k	method	ite	found	NMSE	time
Binary Tree 256×512	70	OMP (OMP-SR)	70	70	$<1e{-}11$	0.1105 (0.0798)
		gOMP (BSR)	25	70	$<1e{-}11$	0.02778 (0.024294)
118 Bus 59×118	14	OMP (OMP-SR)	14	14	$<1e{-}11$	0.0284 (0.0172)
		gOMP (BSR)	5	14	$<1e{-}11$	0.00609 (0.003876)
118 Bus 59×118	100	OMP (OMP-SR)	100	86	0.8406	0.0706 (0.0582)
		OMP (OMP-SR)	118	100	$<1e{-}11$	0.0922 (0.0659)
		gOMP (BSR)	40	100	$<1e{-}11$	0.03400 (0.025842)
1354 Bus 677×1354	270	OMP (OMP-SR)	270	215	0.2029	3.7723 (2.1686)
		OMP (OMP-SR)	336	270	$<1e{-}11$	5.4101 (3.4590)
		gOMP (BSR)	96	270	$<1e{-}11$	1.2641 (0.982119)

Table 5. Results for the phantom image with increasing noise level $\|\boldsymbol{\varepsilon}\|_2$. Reported normalized approximation error $\frac{\|\boldsymbol{y}-\boldsymbol{\Phi x}\|_2}{\|\boldsymbol{y}\|_2}$, and running time in seconds. Running time of our methods is highlighted in blue. OMP-SR is faster than OMP, and BSR is faster than gOMP.

Noise	k	method	ite	found	$\frac{\|\boldsymbol{y}-\boldsymbol{\Phi x}\|_2}{\|\boldsymbol{y}\|_2}$	time
$\|\boldsymbol{\varepsilon}\|_2 = 0.1$	641	OMP (OMP-SR)	641	638	0.0231	88.7496 (70.3730)
		OMP (OMP-SR)	644	641	0.0001	94.6122 (72.1749)
		gOMP (BSR)	81	641	0.0001	11.3051 (9.1572)
$\|\boldsymbol{\varepsilon}\|_2 = 50$	641	OMP (OMP-SR)	641	638	0.0777	87.1276 (70.6804)
		OMP (OMP-SR)	644	641	0.0727	95.4393 (71.5545)
		gOMP (BSR)	81	641	0.0720	11.7429 (9.7602)
$\|\boldsymbol{\varepsilon}\|_2 = 100$	641	OMP (OMP-SR)	641	631	0.1470	88.9356 (72.0881)
		OMP (OMP-SR)	684	641	0.1337	165.5339 (138.7913)
		gOMP (BSR)	101	641	0.1287	16.0534 (13.8060)
$\|\boldsymbol{\varepsilon}\|_2 = 150$	641	OMP (OMP-SR)	641	621	0.2687	89.2221 (70.3730)
		OMP (OMP-SR)	789	641	0.1671	134.5601 (101.7830)
		gOMP (BSR)	197	641	0.1671	27.9475 (25.4323)

5 Discussion and Future Work

OMP has the advantage of simplicity. A greedy algorithm such as OMP is easy to implement but difficult to analyze. This work offered significant performance improvement over the classical OMP and its extension gOMP with theoretical analysis for convergence and approximation error bound. In addition, the proposed changes for OMP come from a principled approach. They work well when combined with other heuristic or ensemble approaches. One possible future work direction is to improve the greedy choice by leveraging the structure in the signal model.

In addition, the minimal ℓ_1 norm solution is the sparsest only when the signal is sparse enough [11]. Therefore, another future work direction is to identify the specific measurement matrix property that drives sparsity during ℓ_1 norm minimization and use that to improve the greedy choice in an iterative procedure.

References

1. Bai, L., Maechler, P., Muehlberghuber, M., Kaeslin, H.: High-speed compressed sensing reconstruction on FPGA using OMP and AMP. In: 2012 19th IEEE International Conference on Electronics, Circuits, and Systems (ICECS 2012), pp. 53–56 (2012). https://doi.org/10.1109/ICECS.2012.6463559
2. Candes, E.J.: The restricted isometry property and its implications for compressed sensing. C.R. Math. **346**(9–10), 589–592 (2008)
3. Candes, E.J., Tao, T.: Near-optimal signal recovery from random projections: universal encoding strategies? IEEE Trans. Inf. Theory **52**(12), 5406–5425 (2006)
4. Chang, L.H., Wu, J.Y.: An improved rip-based performance guarantee for sparse signal recovery via orthogonal matching pursuit. IEEE Trans. Inf. Theory **60**(9), 5702–5715 (2014)
5. Chen, S.S., Donoho, D.L., Saunders, M.A.: Atomic decomposition by basis pursuit. SIAM J. Sci. Comput. **20**(1), 33–61 (1998)
6. Chen, S., Donoho, D.: Basis pursuit. In: Proceedings of 1994 28th Asilomar Conference on Signals, Systems and Computers. vol. 1, pp. 41–44 (1994)
7. Dai, W., Milenkovic, O.: Subspace pursuit for compressive sensing signal reconstruction. IEEE Trans. Inf. Theory **55**(5), 2230–2249 (2009). https://doi.org/10.1109/TIT.2009.2016006
8. Daubechies, I.: Time-frequency localization operators: a geometric phase space approach. IEEE Trans. Inf. Theory **34**(4), 605–612 (1988)
9. Dirksen, S., Lecué, G., Rauhut, H.: On the gap between restricted isometry properties and sparse recovery conditions. IEEE Trans. Inf. Theory **64**(8), 5478–5487 (2018)
10. Donoho, D.L., Tsaig, Y., Drori, I., Starck, J.: Sparse solution of underdetermined systems of linear equations by stagewise orthogonal matching pursuit. IEEE Trans. Inf. Theory **58**(2), 1094–1121 (2012). https://doi.org/10.1109/TIT.2011.2173241
11. Donoho, D.L.: For most large underdetermined systems of linear equations the minimal ℓ_1-norm solution is also the sparsest solution. Commun. Pure Appl. Math. **59**(6), 797–829 (2006)
12. Donoho, D.L., Elad, M.: Optimally sparse representation in general (nonorthogonal) dictionaries via ℓ^1 minimization. Proc. Natl. Acad. Sci. **100**(5), 2197–2202 (2003)

13. Ge, H., Chen, W.: Recovery of signals by a weighted $\ell_2\ell_1$ minimization under arbitrary prior support information. Signal Process. **148**, 288–302 (2018)
14. Huang, G., Wang, L.: High-speed signal reconstruction with orthogonal matching pursuit via matrix inversion bypass. In: 2012 IEEE Workshop on Signal Processing Systems, pp. 191–196. IEEE (2012)
15. Huang, G., Wang, L.: An FPGA-based architecture for high-speed compressed signal reconstruction. ACM Trans. Embed. Comput. Syst. **16**(3), 1–23 (2017). https://doi.org/10.1145/3056481
16. Knoop, B., Rust, J., Schmale, S., Peters-Drolshagen, D., Paul, S.: Rapid digital architecture design of orthogonal matching pursuit. In: 2016 24th European Signal Processing Conference (EUSIPCO), pp. 1857–1861 (2016). https://doi.org/10.1109/EUSIPCO.2016.7760570
17. Kwon, S., Wang, J., Shim, B.: Multipath matching pursuit. IEEE Trans. Inf. Theory **60**(5), 2986–3001 (2014). https://doi.org/10.1109/TIT.2014.2310482
18. Mallat, S., Zhang, Z.: Matching pursuits with time-frequency dictionaries. IEEE Trans. Signal Process. **41**(12), 3397–3415 (1993)
19. Needell, D., Tropp, J.A.: CoSaMP: iterative signal recovery from incomplete and inaccurate samples. Appl. Comput. Harmon. Anal. **26**(3), 301–321 (2009)
20. Needell, D., Vershynin, R.: Uniform uncertainty principle and signal recovery via regularized orthogonal matching pursuit. Found. Comput. Math. **9**(3), 317–334 (2009)
21. Pati, Y., Rezaiifar, R., Krishnaprasad, P.: Orthogonal matching pursuit: recursive function approximation with applications to wavelet decomposition. In: Proceedings of 27th Asilomar Conference on Signals, Systems and Computers. vol. 1, pp. 40–44 (1993)
22. Roy, S., Acharya, D.P., Sahoo, A.K.: Fast OMP algorithm and its FPGA implementation for compressed sensing-based sparse signal acquisition systems. IET Circ. Devices Syst. **15**(6), 511–521 (2021). https://doi.org/10.1049/cds2.12047, https://ietresearch.onlinelibrary.wiley.com/doi/abs/10.1049/cds2.12047
23. Stanislaus, J.L.V.M., Mohsenin, T.: Low-complexity FPGA implementation of compressive sensing reconstruction. In: 2013 International Conference on Computing, Networking and Communications (ICNC), pp. 671–675 (2013). https://doi.org/10.1109/ICCNC.2013.6504167
24. Sturm, B.L., Christensen, M.G.: Comparison of orthogonal matching pursuit implementations. In: 2012 Proceedings of the 20th European Signal Processing Conference (EUSIPCO), pp. 220–224 (2012)
25. Tropp, J.A.: Greed is good: algorithmic results for sparse approximation. IEEE Trans. Inf. Theory **50**(10), 2231–2242 (2004)
26. Tropp, J.A., Gilbert, A.C.: Signal recovery from random measurements via orthogonal matching pursuit. IEEE Trans. Inf. Theory **53**(12), 4655–4666 (2007)
27. Wang, J., Kwon, S., Shim, B.: Generalized orthogonal matching pursuit. IEEE Trans. Signal Process. **60**(12), 6202–6216 (2012). https://doi.org/10.1109/TSP.2012.2218810
28. Yin, P., Lou, Y., He, Q., Xin, J.: Minimization of ℓ_{1-2} for compressed sensing. SIAM J. Sci. Comput. **37**(1), A536–A563 (2015)
29. Yu, Z., et al.: Fast compressive sensing reconstruction algorithm on FPGA using orthogonal matching pursuit. In: 2016 IEEE International Symposium on Circuits and Systems (ISCAS), pp. 249–252 (2016). https://doi.org/10.1109/ISCAS.2016.7527217

MPGTSRN: Scene Text Image Super-Resolution Guided by Multiple Visual-Semantic Prompts

Mingjun Li, Zeming Zhuang, Shuo Xu, and Feng Su(✉)

State Key Laboratory for Novel Software Technology Nanjing University,
163 Xianlin Road, Nanjing, China
{limingjun,zmzhuang,xushuo}@smail.nju.edu.cn, suf@nju.edu.cn

Abstract. Scene text image super-resolution (STISR) aims at enhancing the visual clarity of a low-resolution text image for human perception or tasks like text recognition. In recent STISR work, various visual and semantic clues of the text play a key role in recovering the details of the text, but the utilization of different clues and their interactions is still insufficient, which often results in distorted or blurred appearances of the reconstructed text. To address this problem, we propose a multi-prompt guided text image super-resolution network (MPGTSRN). Specifically, we introduce multiple visual prompts for the text and combine them with semantic features to comprehensively capture the diverse characteristics of the text. We then propose a recurrent reconstruction network integrating multiple visual-semantic prompts to enhance the representation of the text and yield a high-resolution text image. We further propose a cross-representation attention mechanism that utilizes the complementarity of different prompts to guide the reconstruction network to adaptively focus on salient parts of the text and effectively improves the text details. The experimental results show the superiority of our proposed MPGTSRN in the STISR task.

Keywords: Super-resolution · Scene text image · Multiple prompts · Visual-semantic clue

1 Introduction

The textual content in images is an important source of information in people's daily lives. However, scene text images often suffer from various forms of quality degradation such as low resolution and blurring, which hinders the reliable extraction and interpretation of the textual information in the image. For example, most recent deep learning-based text recognition methods do not work well enough on low-resolution images [3,25,31]. Therefore, scene text image super-resolution (STISR), which aims to enhance the resolution and visual clarity of

M. Li and Z. Zhuang—Equal contribution.

A. Antonacopoulos et al. (Eds.): ICPR 2024, LNCS 15332, pp. 316–331, 2025.
https://doi.org/10.1007/978-3-031-78125-4_22

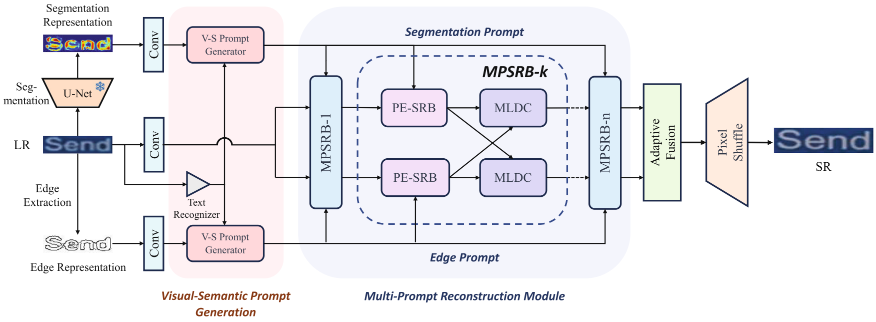


Fig. 1. The architecture of the proposed text image super-resolution network MPGTSRN.

the text in low-resolution images, has a wide range of applications in various text-related fields.

A variety of STISR methods have been proposed in recent years, which can be roughly categorized into two schemes—generic approaches and text-specific approaches. Taking text images as general images, most existing single-image super-resolution methods can be employed for STISR, and variant deep neural network models such as convolutional neural network (CNN) and generative adversarial network (GAN) have been used to learn the LR-HR mapping and accordingly reconstruct the super-resolution (SR) image.

To exploit the characteristics of the text to improve the quality of the reconstructed text image, recent STISR methods [9,15,16,25,27,31] introduce various text clues such as character-level or stroke-level appearance/structural clues and probability-based semantic clues, which capture inherent visual or semantic characteristics of text, to guide the super-resolution process and usually achieve better SR quality and higher text recognition accuracy than general image SR models. For example, TATT [16] introduces text semantic priors into the model and exploits them as guidance for the text reconstruction process. Similarly, C3-STISR [31] introduces visual and linguistic clues of text to help generate higher quality text images.

Despite the encouraging results achieved, existing STISR methods still suffer from loss or distortion of text details such as blurred edges and irregular character shapes in the output text image. The use of a wider range of characteristics of the text in SR models has been shown to be an effective way to improve SR results, but the forms of text cues utilized so far are still quite limited and monolithic, and few studies have focused on modeling and utilizing correlations between different cues to improve their effectiveness in STISR.

In this paper, we propose a novel multi-prompt guided text image super-resolution network MPGTSRN, which introduces and leverages multiple visual-semantic prompts and their interactions to guide the recurrent reconstruction process and effectively improves the quality of the generated high-resolution text images. Figure 1 shows the overall architecture of MPGTSRN.

The main contributions of our work are summarized as follows:

- We introduce edge- and segmentation-based prompts as additional information channels of the text image super-resolution model, which capture distinctive and complementary appearance details of the text and are further combined with semantic features to provide richer clues of the text for the reconstruction model.
- We propose an effective multi-prompt reconstruction network for text images. The network integrates multiple recurrent reconstruction branches to progressively enhance the representation of the text utilizing different text cues captured by each prompt, and finally aggregates the outputs of all branches to yield a high-resolution text image with improved text details.
- We propose a cross-representation attention mechanism to exploit the complementarity of different prompts to guide the multi-prompt reconstruction network to adaptively focus on salient parts of the text in the reconstruction, which effectively enhances the super-resolution results.
- Our method achieves leading performance on the mainstream TextZoom benchmark, demonstrating the effectiveness of the proposed SR model.

2 Related Work

2.1 Single Image Super-Resolution

General single-image super-resolution techniques aim to generate a high-resolution image with recovered details based on its low-resolution (LR) counterpart through learning the mapping from LR patches to HR patches. For this purpose, some methods such as SRCNN [6], EDSR [13] and RDN [28] employ convolutional neural networks to learn the LR-HR mapping in an end-to-end framework. On the other hand, some methods such as SRGAN [11], GLEAN [2] and LDL [12] train generative adversarial networks to recover missing realistic image details for the SR task and improve the restoration quality with diverse losses and priors. Despite their effectiveness on generic images, these methods do not take advantage of the characteristics of text and therefore cannot achieve optimal performance in the STISR task.

2.2 Scene Text Image Super-Resolution

Early scene text image super-resolution methods [19] employed general SR models with varied deep network structures to enhance the resolution of input text images. To further improve the quality of the output text image, most recent STISR methods [3,4,25] introduced and exploited specific characteristics of text in the recovery process. For example, TSRN [25] employs BLSTM to model the sequential characteristic of text and introduces gradient profile loss to help reconstruct high quality text images. Text Gestalt [4] proposes a stroke focused module (SFM) to concentrate more on stroke regions with the guidance of stroke-level attention maps. Besides the structural property, a variety of other text clues have also been utilized. For instance, the semantic features of the text (aka. text prior) are often employed in recent STISR work [15,16,31] as a guidance for text reconstruction, which usually take the form of a vector of character classification probability distributions obtained using an additional text recognizer. C3-STISR [31] further introduces visual and linguistic clues to improve the details of the generated text images. LEMMA [9] introduces an explicit character location modeling mechanism to distinguish character regions from the background.

Our work extends previous clue-guided STISR methods which mostly exploit the features extracted from the text image as the only visual clue, by introducing edge- and segmentation-based prompts and their interactions as additional clues for recovering the text with enhanced readability.

3 Methodology

Our proposed MPGTSRN improves the text image super-resolution results over previous methods through two main mechanisms. First, MPGTSRN introduces edge and segmentation prompts of the text and combines them with semantic cues as effective text clues for the reconstruction model. Second, MPGTSRN employs an effective multi-branch recurrent reconstruction framework with cross-modal attention mechanism to enhance the representation of the text under the guidance of multiple prompts and their complementarity to produce high-quality text images.

As shown in Fig. 1, MPGTSRN consists of five main components: the feature extraction backbone, the visual-semantic (V-S) prompts generator (VSPG), the multi-prompt reconstruction (MPR) module, the adaptive fusion module, and the pixel shuffle layer [23]. Unlike the image-semantic features exploited by previous STISR methods like TSRN and TATT, the proposed VSPG module combines semantic priors of the text with the edge and segmentation features of the input LR text image to generate complementary visual-semantic prompts, which capture multi-aspect visual and semantic characteristics of the text. The MPR module recurrently enhances the representation of the text with visual-semantic prompts through the sequential recurrent block and cross-representation attention mechanism. Finally, the fusion module adaptively combines the dual representation branches, and the pixel shuffle layer reconstructs the SR text image based on the enhanced image representations. We describe each component of the network in the following sections.

3.1 Multiple Representations of Text Image

In order to capture multi-aspect visual cues of the text to help recover the details of the high-resolution text, we extract both the edge and segmentation maps of an input low-resolution (LR) text image $I \in \mathbb{R}^{H \times W \times 3}$ (H and W are the height and width of the image) as complementary representations to the original image to provide useful clues of text shape.

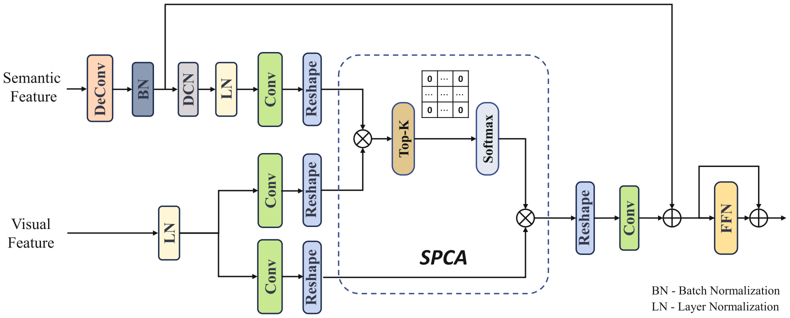


Fig. 2. Visual-semantic prompts generator (VSPG).

Specifically, we employ a pre-trained convolutional network (described in Sect. 4.2) to predict the pixel-level segmentation map of the input LR text image, assigning each pixel a score indicating its probability of belonging to the text or the background. On the other hand, we use the Canny operator to extract the edge map of the LR text image. We then employ a large kernel (9×9) convolution layer to extract the initial feature representations of the input text image and the edge and segmentation maps and capture global, long-range dependencies between the features.

3.2 Visual-Semantic Prompts Generator

The visual-semantic prompts generator integrates the edge and segmentation representations of the text image with semantic features to generate enhanced visual-semantic prompts of the text for text representation reconstruction. Figure 2 shows the structure of the VSPG module.

Specifically, we first feed the input text image to a pre-trained text recognizer [1] to obtain a sequence of character category probability distribution vectors, which are used as semantic features and guidances for relevant locations in the initial edge and segmentation feature representations. Next, as shown in Fig. 2, we apply deconvolution and batch normalization layers on the semantic features to obtain a 2D semantic feature map of the same size as the image representation, and a deformable convolutional network (DCN) is further employed to model the correlations in the feature representation.

To enhance the edge and segmentation representations with the semantic features of the text, inspired by Transformer [24] attention mechanism and the work [30], we propose a sparse cross-attention (SPCA) block to capture and incorporate global correlations between semantic features and the edge/segmentation features to generate *edge prompt* and *segmentation prompt* respectively. The structure of the SPCA block is shown in Fig. 2, which can be formulated as:

$$\mathrm{SPCA}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \mathrm{softmax}\left(\varPhi_{top-k}\left(\frac{\mathbf{Q}\mathbf{K}^{\top}}{\sqrt{d_k}}\right)\right)\mathbf{V} \tag{1}$$

$$[\Phi_{top-k}(S)]_{ij} = \begin{cases} S_{ij}, \, S_{ij} \in \mathrm{TopK}\,(S) \\ 0, \quad otherwise \end{cases} \tag{2}$$

$$\mathbf{X}_q' = \mathbf{X}_q + \mathrm{SPCA}\left(\mathrm{Conv}\left(\mathrm{LN}\left(\mathrm{DCN}\left(\mathbf{X}_q\right)\right)\right), \mathrm{Conv}\left(\mathrm{LN}\left(\mathbf{X}_k\right)\right), \mathrm{Conv}\left(\mathrm{LN}\left(\mathbf{X}_v\right)\right)\right) \quad (3)$$

$$\mathbf{X}_{out} = \mathbf{X}_q' + \mathrm{FFN}\left(\mathbf{X}_q'\right) \quad (4)$$

where the 2D semantic feature map $\mathbf{X}_q$ is taken as $\mathbf{Q}$, and the edge or segmentation feature map is taken as $\mathbf{K}$ and $\mathbf{V}$ (denoted by $\mathbf{X}_k$ and $\mathbf{X}_v$). The similarity matrix S is calculated and sorted in descending order based on the similarity scores between features in $\mathbf{Q}$ and $\mathbf{K}$ respectively, and the operation Φ_{top-k} further retains only the top K scores while ignoring the other scores and corresponding features. We then multiply the matrix resulting from the softmax operation with $\mathbf{V}$ and obtain the final edge/segmentation prompt $\mathbf{X}_{out}$ after the residual connection and the feed-forward network (FFN). In this way, the impact of artifacts in the edge and segmentation maps can be adaptively alleviated by the top-k mechanism.

3.3 Recurrent Multi-prompt Reconstruction Pipeline

Given the semantic-enhanced segmentation and edge prompts which have a size of $C \times H \times W$ (C is the number of channels), MPGTSRN reconstructs text image details with a pipeline composed of a series of multi-prompt sequential recurrent blocks (MPSRBs) as shown in Fig. 1. After the last MPSRB, an adaptive fusion module is employed to combine the outputs of the two reconstruction branches of MPSRB to produce the final representation of the reconstructed image.

Multi-prompt Sequential Recurrent Block. MPSRB takes the two output representations from the previous block as the inputs, and employs two reconstruction branches to enhance the image representation with the segmentation and edge prompts respectively and recover the details of the text.

As shown in Fig. 1, a MPSRB branch first employs a prompt-enhanced sequential recurrent block (PE-SRB) to combine image and prompt information for text representation reconstruction. Figure 3 shows the structure of a PE-SRB. Different from the work [16] which adds directly the input image and prompt features together, PE-SRB combines the two features with an adaptive fusion mechanism, which is formulated as:

$$\mathbf{F}_{out} = \mathbf{F}_{in}^3 + \mathbf{F}_{in}^2 \otimes \mathrm{Sigmoid}\left(\mathbf{W}\mathbf{F}_{in}^1\right) \quad (5)$$

where $\otimes$ denotes Hadamard product, and $\mathbf{W}$ is the learned linear transformation. $\mathbf{F}_{in}$ is the concatenation of the two input feature maps along the channel dimension. The resulting feature maps are projected into three different feature spaces $\mathbf{F}_{in}^1$, $\mathbf{F}_{in}^2$ and $\mathbf{F}_{in}^3$ through convolution, which are then combined into the final output feature $\mathbf{F}_{out}$ by channel attention.

Next, similar to [25], PE-SRB models the sequential relationships in the combined text features through convolution and BLSTM layers and outputs a representation of the size $C \times H \times W$.

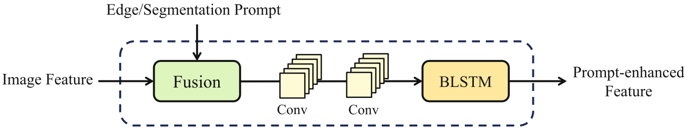


Fig. 3. Prompt-enhanced sequential recurrent block (PE-SRB).

Cross-Representation Attention Through Mutual-Learning Dynamic Convolution. Since the edge and segmentation representations convey rich and complementary appearance details of the text, after integrating the image representations in two reconstruction branches with edge and segmentation prompts respectively, we propose to further make use of the visual clues of the text captured in one representation to guide the enhancement of the other representation through a cross-representation attention (CRA) mechanism.

Specifically, MPGTSRN employs a mutual-learning dynamic convolution (MLDC) block to bridge the information of two reconstruction branches. As shown in Fig. 1, MLDC dynamically predicts the parameters of the convolution that is applied on the representation in one branch n, based on the output representation of PE-SRB in the other branch n'. In this way, a reconstruction branch can utilize the text clues obtained by the other branch to adaptively focus on certain salient parts of the text representation to improve the recovered details of the text.

Moreover, to better capture the linear morphological structure characteristics of character strokes, inspired by the work [20], we employ the dynamic snake convolution (DSConv) in the proposed MLDC block. As shown in Fig. 4, DSConv straightens the standard convolution kernel in both the x- and y-axis and augments each sampling (grid) position of the kernel by a predicted offset. Taking a DSConv kernel of size 9 and the x-axis direction as an example, the specific position of each grid in the kernel K is represented as $K_{i\pm c} = (x_{i\pm c}, y_{i\pm c})$, where $c \in [0, 4]$ represents the horizontal distance from the central grid K_i. The offset of the grid $K_{i\pm c}$ relative to K_i is the summation of the predicted offset between every pair of neighbouring grids from K_i to $K_{i\pm c}$. The kernel also includes a series of sampling grids in the y-axis direction similarly.

Note that, to allow mutual learning between two reconstruction branches, as shown in Fig. 4, we modify the original DSConv model, which uses the same features for the grid prediction and the convolution operation, by taking the features $\mathbf{F}'_{in}$ from the complementary branch as the input for predicting the grids of the DSConv kernel, which is applied on the features $\mathbf{F}_{in}$ of the current branch to produce the output features $\mathbf{F}_{out}$. The experiment results demonstrate the effectiveness of our proposed CRA mechanism based on MLDC and DSConv in improving the recovery quality of the text.

Adaptive Fusion Module. The adaptive fusion (AF) module fuses the complementary representations of the text image obtained by two reconstruction branches with dynamically computed aggregation weights to produce more accurate text details.

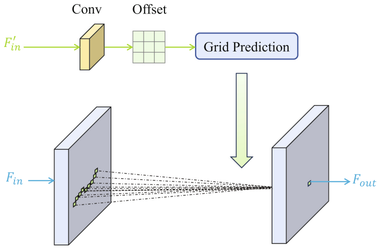


Fig. 4. Mutual-learning dynamic convolution (MLDC) block based on DSConv.

Given the reshaped representations $\mathbf{F}_s \in \mathbb{R}^{C \times N}$ and $\mathbf{F}_e \in \mathbb{R}^{C \times N}$ $(N = H \times W)$ output by the two branches of the last MPSRB, AF first predicts the fusion weights $\mathbf{W}_F$ based on $\mathbf{F}_s$ and $\mathbf{F}_e$, which adaptively adjust the importance of individual features. The fusion weights are then multiplied by the transformed features to obtain the aggregated enhanced representation of the text image. The fusion mechanism can be formulated as follows:

$$\begin{aligned} \mathbf{W}_F &= \text{Sigmoid}\left(\mathbf{W}_1\left[\mathbf{F}_s; \mathbf{F}_e\right]\right) \\ \bar{\mathbf{F}} &= \mathbf{W}_F \odot \left(\mathbf{W}_2\left[\mathbf{F}_s; \mathbf{F}_e\right]\right) \end{aligned} \tag{6}$$

where $\mathbf{W}_1$ and $\mathbf{W}_2$ are learned linear transformations, $\odot$ denotes the element-wise multiplication operation, and $\bar{\mathbf{F}}$ is the fused and enhanced representation of the text image.

Finally, a pixel shuffle layer [23] is employed to transform the enhanced representation $\bar{\mathbf{F}}$ into the output SR text image.

3.4 Loss Function

MPGTSRN is trained with a loss function consisting of the super-resolution loss L_{SR}, the text prior loss L_{TP}, and the text-focused loss L_{TFL} proposed in [3]:

$$L = \lambda_1 L_{SR} + \lambda_2 L_{TP} + \lambda_3 L_{TFL} \tag{7}$$

where λ_1, λ_2, and λ_3 are the balancing weights, which are set to 1.0, 0.01, and 0.5 as in [9].

The super-resolution loss L_{SR} is the L2 norm of the difference between the SR output and the ground-truth high-resolution (HR) image. The text prior loss L_{TP} is the cross-entropy loss of the pre-trained text recognizer used for semantic feature extraction:

$$L_{TP} = \text{Cross-entropy}\left(p_{rec}, y_{label}\right) \tag{8}$$

where p_{rec} denotes the character probability distribution vectors predicted by the pre-trained text recognizer on LR images, and y_{label} denotes the ground truth. A detailed description of the text-focused loss L_{TFL} can be found in [3].

4 Experiments

4.1 Dataset

We evaluate MPGTSRN on the state-of-the-art STISR dataset TextZoom [25], which includes 21,740 LR-HR text image pairs captured through lens zooming in real-world scenarios. The training set consists of 17,367 image pairs, while the test set is further divided into three sets—easy (1,619 samples), medium (1,411 samples), and hard (1,343 samples) based on camera focal length.

4.2 Implementation Details

We implement MPGTSRN based on the PyTorch framework and conduct the experiments on two NVIDIA Tesla V100 GPUs. The multi-prompt reconstruction module comprises five MPSRB blocks, and the text recognizer proposed in [1] is used in MPGTSRN to generate the semantic features of the text. The hyperparameters in our method are consistent with the previous methods TSRN [25] and TATT [16]. We train the network using Adam optimizer with a learning rate of 0.001. LR and HR images are resized to 16×64 and 32×128 respectively.

The text segmentation network mentioned in Sect. 3.1 has an U-Net structure. We generate the ground-truth text segmentation map by applying K-means clustering on image pixels, with K being set to 2 corresponding to the text and background categories. We train the text segmentation network on the synthetic MJSynth (MJ) dataset [10] for one epoch with a learning rate of 0.001 since the image backgrounds in the dataset are simple enough to distinguish from the text foreground. MPGTSRN consists of 29.95M parameters and achieves an average processing time of 3.49ms per image in the inference on the benchmark dataset.

4.3 Ablation Study

Effectiveness of Multiple Prompts in Text Reconstruction. We verify the effectiveness of the proposed multi-prompt framework for text representation enhancement and reconstruction. In Table 1, we compare the text recognition accuracy obtained by the CRNN [21] recognizer on the SR images generated by the proposed model and two variant models which make use of the edge or segmentation prompts solely. It can be seen that the edge prompt is slightly more effective than the segmentation prompt for recovering the text representation for recognition, as it better captures high-frequency shape characteristics of character which play an important role in the recognition. By further incorporating the supplementary text clues provided by the segmentation prompt, the proposed model achieves overall improved recognition accuracies on all benchmarks owing to the higher-quality text image generated based on the enhanced representation.

Table 2 further shows the text recognition accuracy on the SR images obtained by a variant of the proposed reconstruction model that replaces the edge and segmentation prompts with the text semantic prior used in [31], which is composed of a sequence of probability distribution vectors predicted by the text recognizer. Compared to it, the proposed multiple prompts effectively enhance the representation of the text with distinctive visual cues of characters captured by the edge and segmentation maps, which help to enhance the super-resolution and recognition accuracy and demonstrate the effectiveness of the proposed mechanism.

Table 1. Text recognition accuracy (%) of the SR images obtained with variant prompts

Model	Easy	Medium	Hard	Avg
Edge prompt	67.9	60.5	45.9	58.8
Seg. prompt	66.1	58.0	44.7	56.7
Proposed	**68.8**	**60.6**	**46.8**	**59.4**

Table 2. Text recognition accuracy (%) of the SR images obtained with the text semantic prior and the proposed multiple prompts

Model	Easy	Medium	Hard	Avg
Text prior [31]	66.2	59.7	45.5	57.7
Multi-prompts	**68.8**	**60.6**	**46.8**	**59.4**

Table 3. Effectiveness of the sparse cross-attention (SPCA) block

Model	Easy	Medium	Hard	Avg
MHCA	67.6	60.2	45.2	58.3
DSTA [31]	67.6	58.7	**46.8**	58.4
SPCA	**68.8**	**60.6**	**46.8**	**59.4**

Table 4. Effectiveness of the cross-representation attention (CRA) mechanism

Model	Easy	Medium	Hard	Avg
w/o CRA	**69.3**	58.5	44.2	58.1
CRA w. DC	68.8	60.3	45.7	59.0
CRA w. DCN	69.0	60.0	46.5	59.2
CRA w. MLDC	68.8	**60.6**	**46.8**	**59.4**

Effectiveness of Sparse Cross-Attention in Multi-modal Fusion. We compare the proposed sparse cross-attention (SPCA) block with the standard multi-head cross-attention (MHCA) of Transformer and the DSTA block employed in C3-STISR [31] for fusing features of different modalities for prompt generation. As shown in Table 3, compared to MHCA and DSTA which integrates concatenated visual and semantic features through deformable convolution and channel attention, SPCA employs a more effective and flexible query-based attention with top-k filtering to reduce the influence of defective features in fusing multi-modal representations, which helps to improve the recognition accuracy of the reconstructed text.

Figure 5 presents some visualizations of the feature maps resulted from the top-k weighting in SPCA. The highlighted features mostly concentrate in salient positions in the edge and segmentation maps, which shows the effectiveness of the mechanism.

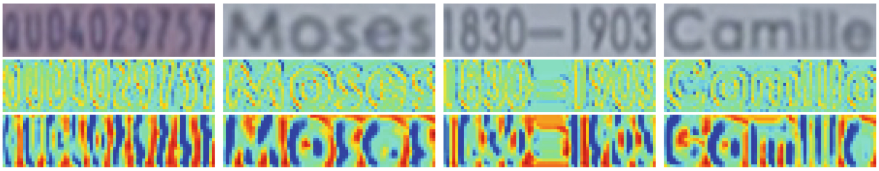


Fig. 5. Examples of the edge-based (2nd row) and segmentation-based (3rd row) feature maps resulted from the top-k weighting in SPCA on some LR images (1st row).

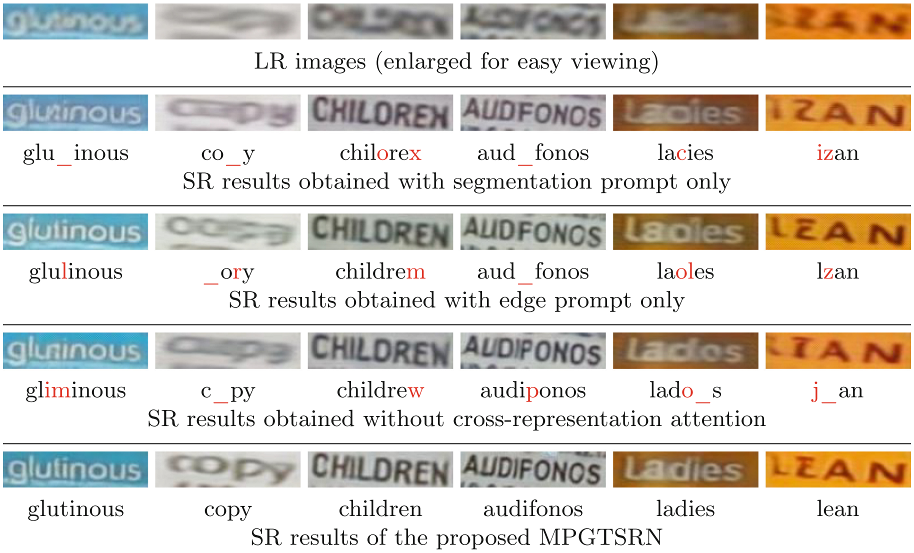

LR images (enlarged for easy viewing)
glu_inous
co_y
chilorex
aud_fonos
lacies
izan
SR results obtained with segmentation prompt only
glulinous
_ory
childrem
aud_fonos
laoles
lzan
SR results obtained with edge prompt only
gliminous
c_py
childrew
audiponos
lado_s
j_an
SR results obtained without cross-representation attention
glutinous
copy
children
audifonos
ladies
lean
SR results of the proposed MPGTSRN

Fig. 6. Examples of the reconstructed SR text images obtained by variant models in the ablation study. The corresponding text recognition results are displayed below the images. Incorrect recognition results are displayed in red text, and '_' denotes a missing character in the recognition result. (Color figure online)

Effectiveness of Cross-Representation Attention. We evaluate the effect of the proposed cross-representation attention (CRA) mechanism on text reconstruction. Table 4 compares the text recognition accuracy on the SR images generated by the SR model without the CRA mechanism and three SR models with different CRA implementations, one using dynamic convolution (DC) with conventional convolution kernels [5], one using deformable convolutional network (DCN), and one using the proposed DSConv-based MLDC, respectively. Both DC and DCN have two feature inputs as the proposed MLDC, one from the current reconstruction branch and the other one from the complementary branch.

Comparing the results of the models with and without the CRA mechanism, we can see that the cross-representation attention effectively improves the recognition accuracy, especially for images with significant loss of textual structural information, by exploiting the knowledge about the target text obtained from the other complementary representation to adaptively enhance the text representation in the current reconstruction branch through mutual-learning dynamic convolution. Moreover, it can be seen that, compared to the DC and DCN variants, the proposed MLDC based on DSConv further enhances the SR results as the introduced constraints on convolution kernel shapes make it easier for the reconstruction module to recover stroke-level information of text and improve the readability of the generated SR images.

Figure 6 presents some examples of the reconstructed SR text images obtained by the proposed MPGTSRN and the variant models that are compared in the ablation study, along with corresponding text recognition results yielded

Table 5. Text recognition accuracy (%) of the SR images obtained by different methods

Method	ASTER [22]				MORAN [14]				CRNN [21]			
	Easy	Med.	Hard	Avg	Easy	Med.	Hard	Avg	Easy	Med.	Hard	Avg
SRCNN [7]	70.6	44.0	31.5	50.0	63.9	40.0	29.4	45.6	41.1	22.3	22.0	29.2
SRResNet [11]	69.4	50.5	35.7	53.0	66.0	47.1	33.4	49.9	45.2	32.6	25.5	35.1
HAN [18]	71.1	52.8	39.0	55.3	67.4	48.5	35.4	51.5	51.6	35.8	29.0	39.6
TSRN [25]	75.1	56.3	40.1	58.3	70.1	53.3	37.9	54.8	52.5	38.2	31.4	41.4
TBSRN [3]	75.7	59.9	41.6	60.0	74.1	57.0	40.8	58.4	59.6	47.1	35.3	48.1
PCAN [4]	77.5	60.7	43.1	61.5	73.7	57.6	41.0	58.5	59.6	45.4	34.8	47.4
TG [29]	77.9	60.2	42.4	61.3	75.8	57.8	41.4	59.4	61.2	47.6	35.5	48.9
TPGSR [15]	77.0	60.9	42.4	60.9	72.2	57.8	41.3	57.8	61.0	49.9	36.7	49.8
TATT [16]	78.9	63.4	45.4	63.6	72.5	60.2	43.1	59.5	62.6	53.4	39.8	52.6
C3-STISR [31]	79.1	63.3	46.8	64.1	74.2	61.0	43.2	60.5	65.2	53.6	39.8	53.7
TSAN [33]	79.6	64.1	45.3	64.1	78.4	61.3	45.1	62.7	64.6	53.3	38.8	53.0
LEMMA [9]	81.1	66.3	47.4	66.0	77.7	62.5	44.6	63.2	67.1	58.8	40.6	56.3
RGDiffSR [32]	81.1	65.4	49.1	66.2	78.6	62.1	45.4	63.1	67.6	56.5	42.7	56.4
RTSRN [27]	80.4	66.1	49.1	66.2	77.1	63.3	46.5	63.2	67.0	59.2	42.6	57.0
MPGTSRN	**82.1**	**68.9**	**52.6**	**68.8**	**80.2**	**66.3**	**50.3**	**66.5**	**68.8**	**60.6**	**46.8**	**59.4**

by the CRNN [21] text recognizer. Comparing the images obtained with the edge prompt and the segmentation prompt respectively, e.g. the first two examples, we can see that both prompts have some advantages in handling variant degraded text and are usually complementary to each other. By integrating the visual clues of the text captured by both prompts, MPGTSRN effectively overcomes the defects in text segmentation and edge extraction caused by blur and distortion and accurately recovers the details of the text. On the other hand, as cross-representation attention provides an effective mechanism for joint enhancement of two complementary representations, MPGTSRN significantly improves the quality of the recovered text image compared to the model without CRA, as shown in all the examples in Fig. 6.

4.4 Comparison with State-of-the-Arts

We compare the text recognition accuracy of the SR images obtained by MPGTSRN and some other state-of-the-art STISR methods that employ the same training data and settings as ours in Table 5. As common practice, three different pre-trained text recognition models, ASTER [22], MORAN [14], and CRNN [21] which are not fine-tuned on the TextZoom STISR dataset, are employed to recognize the SR text image. For the fairness of comparison, we do not include some methods that additionally trained these test recognizers end to end on the SR dataset in the comparison.

Table 6. Performance comparison using ABINet, MATRN and PARSeq recognizers on SR images obtained by different methods

Method	ABINet [8]				MATRN [17]				PARSeq [1]			
	Easy	Med.	Hard	Avg	Easy	Med.	Hard	Avg	Easy	Med.	Hard	Avg
TATT [16]	80.7	65.8	50.3	66.5	81.1	66.6	51.7	67.4	82.2	65.9	52.1	67.7
C3-STISR [31]	81.4	66.9	49.9	67.0	81.9	68.0	51.1	68.0	84.3	68.3	50.9	68.8
LEMMA [9]	82.6	69.2	50.6	68.5	82.8	70.4	51.7	69.3	83.6	69.2	52.3	69.3
RGDiffSR [32]	84.6	66.3	53.3	69.1	84.7	67.4	53.5	69.5	84.2	67.5	53.2	69.3
MPGTSRN	**84.8**	**71.1**	**56.1**	**71.6**	**85.0**	**71.3**	**56.7**	**71.9**	**86.0**	**72.0**	**57.5**	**72.7**

MPGTSRN outperforms previous STISR methods in the comparison in all test benchmarks, showing the effectiveness of the proposed multi-prompt guidance mechanism in the STISR task. Particularly, compared to C3-STISR which exploits the language model for generating semantic clues and uses a text skeleton painter to generate additional visual clues of the text, our MPGTSRN makes use of the edge and segmentation clues of the text which can be easily obtained from the standard SR dataset itself and achieves significantly enhanced performance. On the other hand, compared to RTSRN which conducts a multi-stage training strategy to improve the SR performance, MPGTSRN has only gone through one stage of training but still achieves better results than RTSRN.

Table 6 further shows the recognition results obtained using three newer text recognizers ABINet [8], MATRN [17] and PARSeq [1] on the SR results of different methods. MPGTSRN achieves the best results in all comparisons.

In Table 7, we compare the peak signal to noise ratio (PSNR) and structural similarity index measure (SSIM) [26] between the ground-truth HR image and the SR image obtained by MPGTSRN and some representative STISR methods. MPGTSRN achieves PSNR and SSIM that are comparable to those of the best methods. It should be noted that, due to the semantic nature of the text image, the recognition accuracy is usually a more meaningful measurement for comparing different STISR methods than PSNR and SSIM.

Figure 7 shows several super-resolution results obtained by some representative STISR models and the corresponding text recognition results yielded by the CRNN recognizer. MPGTSRN demonstrates a remarkable ability to generate text with more regular character shapes and sharper edges, which effectively distinguish the text from the background and in turn lead to more accurate recognition results.

Table 7. PSNR and SSIM

Method	PSNR	SSIM
TSRN [25]	21.4	0.7690
TBSRN [3]	20.9	0.7603
TPGSR [15]	21.0	0.7719
TATT [16]	**21.5**	**0.7930**
C3-STISR [31]	19.8	0.7408
LEMMA [9]	20.9	0.7792
RGDiffSR [32]	21.3	0.7865
MPGTSRN	21.1	0.7788

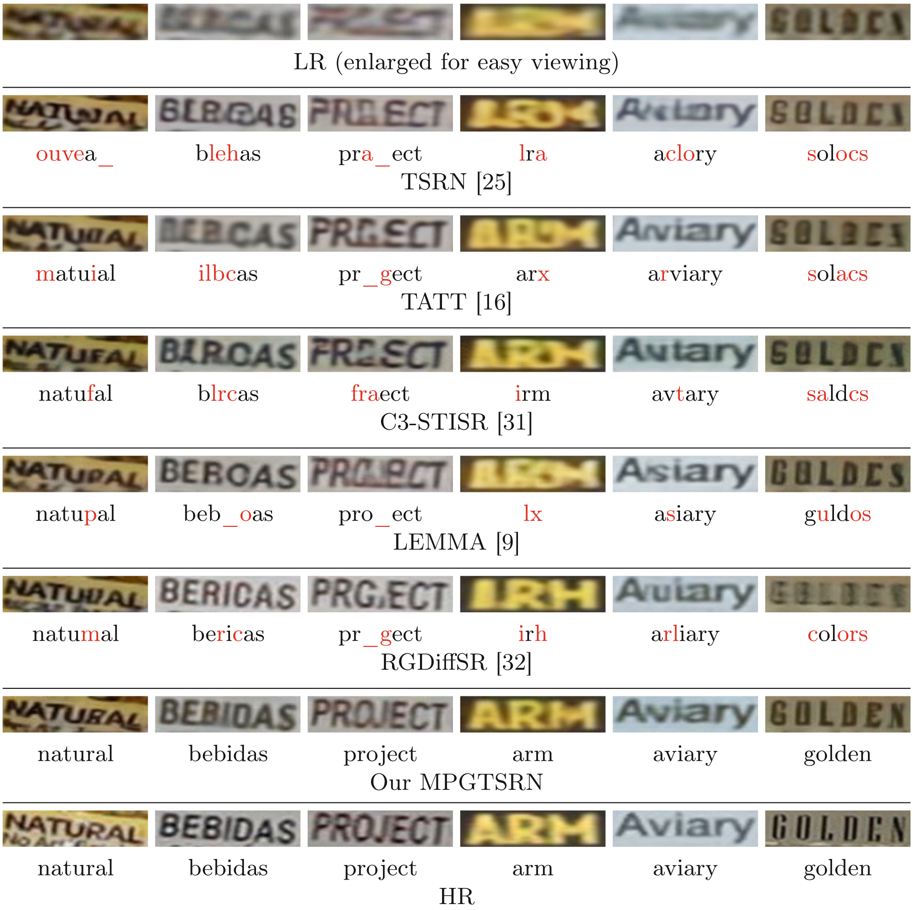

LR (enlarged for easy viewing)
ouvea_ blehas pra_ect lra aclory solocs
TSRN [25]
matuial ilbcas pr_gect arx arviary solacs
TATT [16]
natufal blrcas fraect irm avtary saldcs
C3-STISR [31]
natupal beb_oas pro_ect lx asiary guldos
LEMMA [9]
natumal bericas pr_gect irh arliary colors
RGDiffSR [32]
natural bebidas project arm aviary golden
Our MPGTSRN
natural bebidas project arm aviary golden
HR

Fig. 7. Examples of super-resolution results obtained by different methods. LR denotes the input low-resolution image. HR denotes the ground-truth high-resolution image. Text under an image is the recognition result, in which the red characters are incorrectly recognized ones and '_' denotes a missing character. (Color figure online)

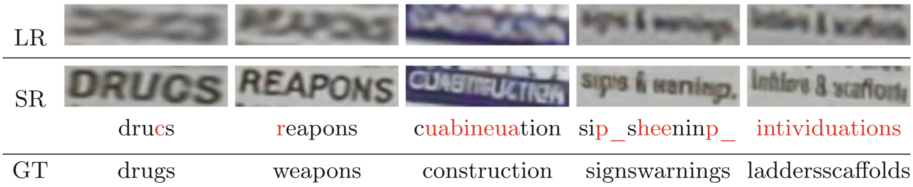

LR
SR
drucs reapons cuabineuation sip_sheeninp_ intividuations
GT
drugs weapons construction signswarnings laddersscaffolds

Fig. 8. Examples of the failure cases of MPGTSRN. Text under an image is the recognition result, in which the red characters are incorrectly recognized ones. 'GT' denotes the ground truth text. (Color figure online)

4.5 Limitations

Although the proposed MPGTSRN demonstrates a good ability to improve the visual clarity of low-resolution text, it's still possible for MPGTSRN to produce incorrectly reconstructed text when dealing with some challenging scene text images. Figure 8 shows some of the failure cases of MPGTSRN, which are usually caused by the very low quality of the input text image such as heavily blurred characters and the severely distorted shape of the text.

5 Conclusions

We present a novel super-resolution network MPGTSRN for scene text images. MPGTSRN introduces edge- and segmentation-based prompts and integrates them with semantic features as effective text clues to enhance the representation of the text. MPGTSRN further introduces dynamic cross-representation attention mechanism to exploit the complementarity of prompts to guide the reconstruction model to yield high-resolution text images with enhanced details and clarity, which effectively improves the accuracy of the subsequent text recognition task.

References

1. Bautista, D., Atienza, R.: Scene text recognition with permuted autoregressive sequence models. In: EECV, pp. 178–196. Springer (2022)
2. Chan, K.C., Wang, X., Xu, X., Gu, J., Loy, C.C.: GLEAN: generative latent bank for large-factor image super-resolution. In: CVPR, pp. 14240–14249 (2021)
3. Chen, J., Li, B., Xue, X.: Scene text telescope: text-focused scene image super-resolution. In: CVPR, pp. 12021–12030 (2021)
4. Chen, J., Yu, H., Ma, J., Li, B., Xue, X.: Text gestalt: stroke-aware scene text image super-resolution. In: AAAI. vol. 36, pp. 285–293 (2022)
5. Chen, Y., Dai, X., Liu, M., Chen, D., Yuan, L., Liu, Z.: Dynamic convolution: attention over convolution kernels. In: CVPR, pp. 11030–11039 (2020)
6. Dong, C., Chen, C.L., He, K., Tang, X.: Learning a deep convolutional network for image super-resolution. In: ECCV, pp. 184–199 (2014)
7. Dong, C., Loy, C.C., He, K., Tang, X.: Image super-resolution using deep convolutional networks. IEEE TPAMI **38**(2), 295–307 (2016)
8. Fang, S., Xie, H., Wang, Y., Mao, Z., Zhang, Y.: Read like humans: autonomous, bidirectional and iterative language modeling for scene text recognition. In: CVPR, pp. 7098–7107 (2021)
9. Guo, H., Dai, T., Meng, G., Xia, S.T.: Towards robust scene text image super-resolution via explicit location enhancement. In: IJCAI, pp. 782–790 (2023)
10. Jaderberg, M., Simonyan, K., Vedaldi, A., Zisserman, A.: Reading text in the wild with convolutional neural networks. IJCV **116**(1), 1–20 (2016)
11. Ledig, C., et al.: Photo-realistic single image super-resolution using a generative adversarial network. In: CVPR, pp. 4681–4690 (2017)
12. Liang, J., Zeng, H., Zhang, L.: Details or artifacts: a locally discriminative learning approach to realistic image super-resolution. In: CVPR, pp. 5647–5656 (2022)

13. Lim, B., Son, S., Kim, H., Nah, S., Lee, K.M.: Enhanced deep residual networks for single image super-resolution. In: CVPRW, pp. 1132–1140 (2017)
14. Luo, C., Jin, L., Sun, Z.: MORAN: a multi-object rectified attention network for scene text recognition. PR **90**, 109–118 (2019)
15. Ma, J., Guo, S., Zhang, L.: Text prior guided scene text image super-resolution. TIP **32**, 1341–1353 (2023)
16. Ma, J., Liang, Z., Zhang, L.: A text attention network for spatial deformation robust scene text image super-resolution. In: CVPR, pp. 5901–5910 (2022)
17. Na, B., Kim, Y., Park, S.: Multi-modal text recognition networks: interactive enhancements between visual and semantic features. In: ECCV, pp. 446–463. Springer (2022)
18. Niu, B., et al.: Single image super-resolution via a holistic attention network. In: ECCV, pp. 191–207. Springer (2020)
19. Peyrard, C., Baccouche, M., Mamalet, F., Garcia, C.: ICDAR2015 competition on text image super-resolution. In: ICDAR, pp. 1201–1205 (2015)
20. Qi, Y., He, Y., Qi, X., Zhang, Y., Yang, G.: Dynamic snake convolution based on topological geometric constraints for tubular structure segmentation. In: ICCV, pp. 6070–6079 (2023)
21. Shi, B., Bai, X., Yao, C.: An end-to-end trainable neural network for image-based sequence recognition and its application to scene text recognition. IEEE TPAMI **39**(11), 2298–2304 (2016)
22. Shi, B., Yang, M., Wang, X., Lyu, P., Yao, C., Bai, X.: ASTER: an attentional scene text recognizer with flexible rectification. IEEE TPAMI **41**(9), 2035–2048 (2018)
23. Shi, W., et al.: Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. In: CVPR, pp. 1874–1883 (2016)
24. Vaswani, A., et al.: Attention is all you need. In: NeurIPS, vol. 30 (2017)
25. Wang, W., et al.: Scene text image super-resolution in the wild. In: ECCV, pp. 650–666. Springer International Publishing, Cham (2020)
26. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: from error visibility to structural similarity. TIP **13**(4), 600–612 (2004)
27. Zhang, W., et al.: Pixel adapter: a graph-based post-processing approach for scene text image super-resolution. In: ACM MM, pp. 2168–2179 (2023)
28. Zhang, Y., Tian, Y., Kong, Y., Zhong, B., Fu, Y.: Residual dense network for image super-resolution. In: CVPR, pp. 2472–2481 (2018)
29. Zhao, C., et al.: Scene text image super-resolution via parallelly contextual attention network. In: ACM MM, pp. 2908–2917 (2021)
30. Zhao, G., Lin, J., Zhang, Z., Ren, X., Su, Q., Sun, X.: Explicit sparse transformer: Concentrated attention through explicit selection (2019)
31. Zhao, M., Wang, M., Bai, F., Li, B., Wang, J., Zhou, S.: C3-STISR: scene text image super-resolution with triple clues. In: IJCAI, pp. 1707–1713 (2022)
32. Zhou, Y., Gao, L., Tang, Z., Wei, B.: Recognition-guided diffusion model for scene text image super-resolution. In: ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 2940–2944 (2024). https://doi.org/10.1109/ICASSP48485.2024.10447585
33. Zhu, X., Guo, K., Fang, H., Ding, R., Wu, Z., Schaefer, G.: Gradient-based graph attention for scene text image super-resolution. In: AAAI. vol. 37, pp. 3861–3869 (2023)

Connecting the Dots: Isolated Trails of Detected Narrow Rivers in Multispectral Images

Jit Mukherjee[1] and Jean-Baptiste Courbot[2](✉)

[1] Department of Computer Science and Engineering, Birla Institute of Technology, Mesra, Ranchi, India
[2] IRIMAS UR 7499 Université de Haute Alsace, Mulhouse, France
jean-baptiste.courbot@uha.fr

Abstract. In this paper, we address the problem of localization of narrow rivers in remote sensing images. Because these rivers may be occluded, thin, or under-resolved, pixel-based methods might not be stable enough to ensure satisfying recovery. In this paper, we propose a two-step approach: first, we detect the main river course and larger segment through a pixel-based approach relying on the normalized difference water index. Second, after missing segments are identified, we propose to connect the dots through a Bézier curve adjustment using a dedicated greedy optimization approach. Results on synthetic and real images show the interest of the proposed approach, with respect to dedicated pixel-based alternatives.

Keywords: Narrow Rivers · Isolated Trails · Curve Optimization

1 Introduction

The monitoring of water bodies is crucial for society, ranging from protecting ecosystems to managing resources. Classical monitoring techniques of river health and networks have been well-documented. However, with the significant technological advancements in electronics and, subsequently, high-performance satellites, the monitoring of land class features from remotely sensed images has become a wide research domain. It includes but not limited to monitoring of water bodies, different fires, forest degradation, and agriculture areas [1]. The contemporary issues of river networks, such as climate change, reduced flow, pollution, and others, necessitate the utilization of advanced and cost-effective technologies [2]. Consequently, fluvial remote sensing has garnered considerable attention [3].

The configuration of surface topography significantly influences the development of river networks, leading to widespread utilization of digital terrain

J. Mukherjee—The author would like to acknowledge support from the French Institute in India through SSHN - SRTF fellowship.

A. Antonacopoulos et al. (Eds.): ICPR 2024, LNCS 15332, pp. 332–341, 2025.
https://doi.org/10.1007/978-3-031-78125-4_23

model (DTM) data in scholarly research for river network detection [4]. The DTM data has proven to be a convenient tool for delineating drainage networks at both global and regional scales [4,5]. However, in hydrological applications, the significant drawback of DTM has found to be its limited accuracy and uncertainties [6,7]. Recognition of river drainage patterns and detection of the delineation of the rivers trails have been few of the domain of interests in Synthetic aperture radar (SAR) image applications [8]. Global water bodies including rivers are mapped through the probability distribution of backscatter - incidence angle combination in [9]. River drainage patterns and delineations are considered in [8,10] through supervised image classifications techniques. However, rivers have heterogeneous features including their wide varieties of width, flow pattern, sedimentation and other. It is difficult to achieve wider applicability by such supervised techniques for a such dynamic land class. Although, there are a few techniques to detect a rivers, the multi-scale characteristics of rivers, mostly narrow rivers, have garnered comparatively less experimentation.

Detection of open water features from multi-spectral images predominantly uses dedicated indexes to enhances features. Several water body indexes such as normalized difference water index (NDWI) [11], modified normalized difference water index (MNDWI) [12], automated water extraction index (AWEI) [13], etc. have been proposed in the literature extensively. However, these indexes operate at a pixel level, and may suffer from mixing different land classes or low contrast, thus may detect narrow rivers as a series of isolated trail segments rather than a continuous one [2,14]. Additionally, separating rivers from other water bodies using such indexes needs more experimentation [2]. Gabor filtering has been found instrumental to enhance objects features in a lower contrast environment, which can be applied to rivers too [15]. However, [15] detects rivers with significant width and forgoes the additional complexities of narrow rivers and their disconnected curvilinear feature. [16] proposes a technique to detect narrow rivers with the presence of highways which shows similar characteristics in spectral domain. They consider the high curvilinear feature as a distinguishing factor, which may not be found useful for a river with significant width. Thus, rivers with significant width and narrow width need to be identified separately, as in [2,17]. However, the authors use a Otsu thresholding, hence assuming a bimodal image histogram, for separating rivers and does not consider the presence of road networks nearby. In [2] a connected-component based technique detects rivers with significant width and narrow width separately. As narrow rivers have lower contrast with the background [14,17] and may suffer from discontinuity, a technique is required where isolated trails of these narrow rivers can be modeled and be connected.

Such identification of river trails primarily requires analysis of linear and curvilinear features, a research problem extensively explored within the field of image processing. This has been explored through various techniques, such as learning-based technique [18], graph based approaches [19], image derivatives [20], gradient vector flow [21]. Most of these methods work on pixel level, including advanced morphology operators: in [22] path opening using $2D$ filtering has been proposed for such delineation, and in [23] a novel operator based on push-pull inhibition is used for the detection of curvilinear structures.

In this paper, we propose to model the problem off-the-grid, *i.e.* to model the curve to search for as a continuous-valued object. The objective is to enforce continuity of the recovered river trails.

2 Methodology

2.1 From a Pixel Perspective

In this part, rivers are detceted in two phases as discussed in [2]. First, rivers with significant widths are detected. Further, narrow rivers are isolated from other open water bodies. Prior to this, water bodies are sensed using NDWI ($\frac{\lambda_{Green} - \lambda_{NIR}}{\lambda_{Green} + \lambda_{NIR}}$). Here λ_{Green} and λ_{NIR} represent reflectance values in green and near infra-red bands. Higher values of the NDWI index highlight open water features, allowing for the detection of significant portions of wide rivers using NDWI. Only isolated segments of narrow rivers are detected by NDWI due to the presence of other land classes such as river sandbank. A lower threshold of NDWI may also detect narrow rivers but will also yield irrelevant regions. It also provides low contrast with the background, hence a Gabor filter based enhancement has been used [2]. Figure 1 summarizes the process for this first part. Nonetheless, numerous portions of narrow rivers remain indiscernible, and endpoints of these trails will be used in the next part.

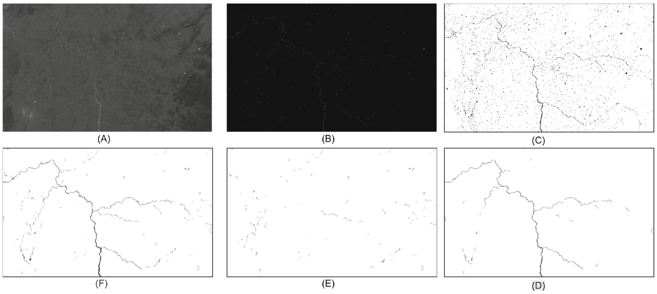


Fig. 1. Pixelwise Technique to Detect Rivers: (A) NDWI Image. (B) Outcome of Gabor Filter. (C) Enhanced Image. (D) Detecting Rivers with Significant Width. (E) Detecting Narrow Rivers. (F) Detected Rivers with isolated trails

2.2 Connecting the Dots

In this section, we provide a method to recover parameterized curves within images, given two known endpoints.

Model and Estimation Problem. We model the curves to recover as Bézier curves with a Gaussian profile. Hence, a given curve is parametrized by K knots (or *control points*) such that $\forall t \in [0,1]$:

$$\mathbf{b}(t) = \sum_{k=0}^{K} b_{k,K}(t)\mathbf{p}_k, \tag{1}$$

with $b_{k,K}$ the Bernstein basis polynomials of degree K, $\mathbf{b}(t)$ and $\mathbf{p}_k \in \mathbb{R}^2$. The curvilinear shape is then transcribed at pixel $\mathbf{s}$ in the image as:

$$x_{\mathbf{s}} = w \exp(-\frac{\|\mathbf{s}-\mathbf{b}\|^2}{2\sigma^2}) \text{ with } \|\mathbf{s}-\mathbf{b}\| = \underset{t\in[0,1]}{\arg\min}\ \|\mathbf{s}-\mathbf{b}(t)\|_2 \tag{2}$$

with $w > 0$ and $\sigma > 0$ the weight and width of the resulting shape with a Gaussian profile.

Hence, given the number of knots K and endpoints $\mathbf{p}_0$ and $\mathbf{p}_K$, estimating the shape of the curve within a given image amounts to finding the adequate w, σ, and intermediate knots $\mathbf{p}_1, \ldots, \mathbf{p}_{K-1}$. In the following, we will denote $\boldsymbol{\theta} = \{w, \sigma, \mathbf{p}_1, \ldots, \mathbf{p}_{K-1}\}$.

Then, given an observation $\mathbf{y}$, the estimation problem for K knots is:

$$\boldsymbol{\theta}^* = \underset{\boldsymbol{\theta}\in\mathbb{R}^+\times\mathbb{R}^+\times\mathbb{R}^{K-2}}{\arg\min}\ \|\mathbf{y}-\mathbf{x}(\boldsymbol{\theta})\|^2, \tag{$\mathcal{P}_K$}$$

with $\mathbf{x}(\boldsymbol{\theta}) \in \mathbb{R}^S$ the image produced from $\boldsymbol{\theta}$ according to Eq. (2).

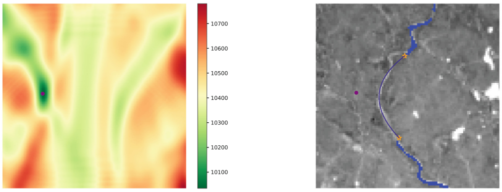


Fig. 2. Criterion of $(\mathcal{P}_3)$ for fixed w, σ and varying position of $\mathbf{p}_1$ (left) and solution at the global minimum (right).

Inverse Problem Considerations. The criterion in $(\mathcal{P}_K)$ is not convex, and more generally present several local minima, as shown in the minimalist example of $\mathcal{P}_3$ in Fig. 2. Noteworthy, local minima are visually not relevant for our purpose. Besides, the number of knots K controls the complexity of the curve, and there is no straightforward method to estimate it beforehand. We propose to take a greedy approach to solve, for a given target K, the intermediates series of problems $\mathcal{P}_3, \ldots, \mathcal{P}_{K-1}$.

Algorithm 1 Greedy curve optimization procedure

```
Require: Number of knots $K > 3$, input NDWI image $\mathbf{y}_0$, endpoints $\mathbf{p}_0$ and $\mathbf{p}_K$
Ensure: Estimation of $\hat{\boldsymbol{\theta}}^{[K]}$
  Preprocessing of $\mathbf{y}_0$, yielding $\mathbf{y}$.
  Initialization. Starting from a grid search on $\mathbf{p}_1$, solve $(\mathcal{P}_3)$ using MH and L-BFGS-B.
  This yields $\hat{\boldsymbol{\theta}}^{[3]}$
  for each $k \leq K$ do
      From $\hat{\boldsymbol{\theta}}^{[k-1]}$, find the closest starting point containing one more knot, via least
  squares.
      Solve $(\mathcal{P}_k)$ using MH and L-BFGS-B, yielding $\hat{\boldsymbol{\theta}}^{[k]}$.
```

Initialization. To avoid local minima, the starting point of the search for a solution to $(\mathcal{P}_3)$ is performed through grid search for the $\mathbf{p}_1$ parameter (see Fig. 2). Thus, the resulting Bézier curve should be reasonably well located in the image.

Pre-processing. To bring closer the image model (2) and the NDWI images, the latter need to be preprocessed. As rivers are expected to be thin, the low frequency in the images are suppressed, and the directional features are enhanced through Gabor filtering.

Solving a $(\mathcal{P}_K)$ *problem.* To yield more chances to reach the global minimum, we split this step in two:

- A coarse optimization, that is made through Metropolis-Hastings (MH) [24] sampling. It is designed to attain more favorable regions which are potentially distant, in the parameter space, from the current point.
- A refined optimization to reach the local minima within the region, and is in practice performed via the L-BFGS-B [25] method.

The overall procedure is summarized in Algorithm 1, and insights on the main intermediate steps are given in Fig 3.

3 Numerical Results

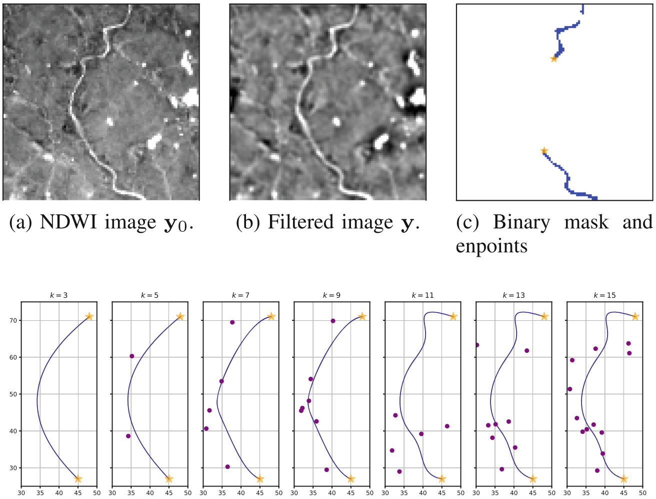
3.1 Synthetic Images

At first, we study the behavior of the proposed method on synthetic images. To generate the latter, we first generate images $\mathbf{x}$ from random curves with varying width (see Fig. 4a) before adding a noisy background $\mathbf{b}$. This background itself is sampled along a Gaussian fractional random field [26], in order to sample uniformly along frequencies. Thus, we generate:

$$\mathbf{y} = \mathbf{x} + \sigma_b \mathbf{b}, \tag{3}$$

where σ_b is tuned according to the signal-to-noise ratio (SNR), defined as:

$$\text{SNR} = 20 \log \left(\frac{\|\mathbf{x}\|_2}{\sigma_b \|\mathbf{b}\|_2} \right). \tag{4}$$



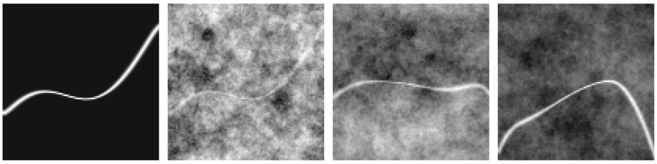
(a) NDWI image $\mathbf{y}_0$.

(b) Filtered image $\mathbf{y}$.

(c) Binary mask and enpoints

(d) Intermediate curves obtained in Alg. 1. Fixed endpoints and control point are depicted in orange and purple respectively.

Fig. 3. Summary of the proposed method. The final result is summarized in Fig. 6g.



(a) $\mathbf{x}$.

(b) $\mathbf{y}$ at $-15\,\text{dB}$.

(c) $\mathbf{y}$ at $-9\,\text{dB}$.

(d) $\mathbf{y}$ at $0\,\text{dB}$.

Fig. 4. Examples of synthetic images.

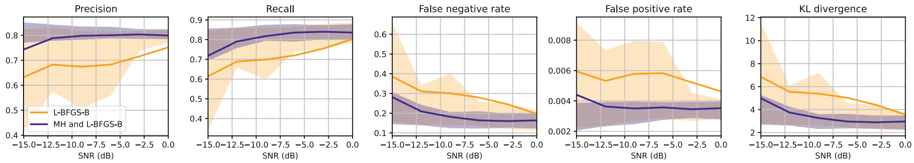


Fig. 5. Average scores on synthetic images. Each point depict the average results over 50 trials, with the shaded regions depicting the first and fourth quartiles.

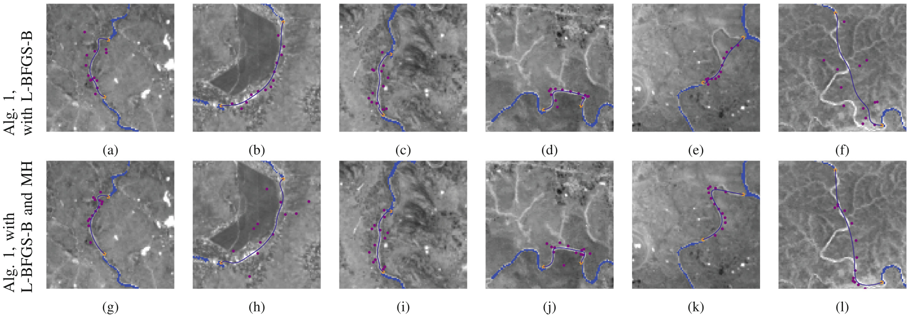


Fig. 6. Example results on real images, with the same legend as in Fig. 3.

The process is depicted in Figs. 4 (b-d).

Sampling several $\mathbf{y}$ under varying SNR, we measure the performance of the considered method (target $K = 15$) in terms of precision, recall, false and true positive rate. KL divergence is additionally calculated to assess the precision of detecting similar curves. Here, the curve is treated as the probability mass and background values are treated as the complement probability. These are used to compute KL divergence values between the ground truth and detected curves. To do so, resulting continuous images are thresholded at $w/10$. Besides, we also evaluate a simpler version of Algorithm 1, for which optimization is made without the coarse MH step. Results are reported in Fig. 5 and suggest the following observations:

- The MH step does effect favorably the results, and in particular at low SNR.
- Overall, the false positive rate is very low. Indeed, the proposed method is conservative, as almost all detected pixel are true positives.
- Our proposed method achieves stable performances between -10 and 0 dB, which is appealing for applications on real images.

3.2 Real Images

Table 1. Result summary on the real images, in percent (excepting KL divergence).

	Error	Prec.	Recall	FP	FN	KL
RORPO [22]	2.3	55.0	32.5	0.7	67.5	11.8
Alg. 1 with L-BFGS-B	2.4	51.2	40.8	0.8	59.2	10.3
Alg. 1 with L-BFGS-B and MH	1.9	64.2	54.2	0.6	45.8	8.0

To evaluate Algorithm 1 on real images, we select by hand regions from the pixel-wise method [2] for which a segment is missing, and locate the endpoints manually. Ground truths were also manually obtained, resulting in a total of 20 real test images. The results on these images are summarized in Table 1 and some examples are given in Fig. 6.

We compare our results with the RORPO [22] method, that aims at finding oriented thin structures in images through path operators. The threshold on the resulting intensity was selected to best match the result of Algorithm 1.

The outcomes can be summarized as follows:

- The RORPO method, while yielding overall interesting results, is outperformed by our proposition. The main explanation is that it is not designed, as most pixel-based method, to yield continuous trails.
- Overall, the addition of the MH step is valuable as well, as all scores are improved.
- Nevertheless, real images remain challenging: indeed, the attained precision and recall are at the same scale as worst-case SNR in synthetic images. This might be explained by the complexity of river courses, as exemplified by abrupt changes seen in Fig 6l. The number of control points, in such case, might also be too low.
- Measures are made pixel-wise, hence a discrepancy might happen because the ground truth have a variable width while our estimation do not.

4 Discussion

Results on real images obtained by our proposition are overall satisfying but suggest also several leads for improvement. First, the number of knots is a parameter that ought to be automatically selected, *e.g.* through an automatic sparsity-aware approach [27]. Besides, there is a need for a global method incorporating segment detection, endpoint detection, and endpoint connection, leveraging the problem to another amplitude, as this implies the recovery of a whole fluvial network. From the viewpoint of remote sensing, linking nearby isolated trails can be addressed using morphological operators. However, connecting distant isolated trails with high sinuosity poses a greater challenge. Mathematical models to represent the sinuosity of narrow rivers and connecting their isolated trails, which this works studies, will be advantageous to different fluvial applications in future.

References

1. Thyagharajan, K.K., Vignesh, T.: Soft computing techniques for land use and land cover monitoring with multispectral remote sensing images: a review. Arch. Comput. Methods Eng. **26**(2), 275–301 (2019)
2. Mukherjee, J.: Identifying rivers with varying width through NDWI from Landsat 8 images. In: IGARSS 2023-2023 IEEE International Geoscience and Remote Sensing Symposium, pp. 3734–3737. IEEE (2023)

3. Andrew Marcus, W., Fonstad, M.A.: Remote sensing of rivers: the emergence of a subdiscipline in the river sciences. Earth Surface Process. Land. **35**(15), 1867–1872 (2010)
4. Pavelsky, T.M., et a.: Assessing the potential global extent of SWOT river discharge observations. J. Hydrol. **519**, 1516–1525 (2014)
5. Liu, Z., Khan, U., Sharma, A.: A new method for verification of delineated channel networks. Water Resour. Res. **50**(3), 2164–2175 (2014)
6. Wechsler, S.P.: Uncertainties associated with digital elevation models for hydrologic applications: a review. Hydrol. Earth Syst. Sci. **11**(4), 1481–1500 (2007)
7. Polidori, L., El Hage, M.: Digital elevation model quality assessment methods: a critical review. Remote Sens. **12**(21), 3522 (2020)
8. Güneralp, İ, Filippi, A.M., Hales, B.U.: River-flow boundary delineation from digital aerial photography and ancillary images using support vector machines. GISci. Remote Sens. **50**(1), 1–25 (2013)
9. Westerhoff, R.S., Kleuskens, M.P.H., Winsemius, H.C., Huizinga, H.J., Brakenridge, G.R., Bishop, C.: Automated global water mapping based on wide-swath orbital synthetic-aperture radar. Hydrol. Earth Syst. Sci. **17**(2), 651–663 (2013)
10. Klemenjak, S., Waske, B., Valero, S., Chanussot, J.: Automatic detection of rivers in high-resolution SAR data. IEEE J. Sel. Top. Appl. Earth Observations Remote Sens. **5**(5), 1364–1372 (2012)
11. McFeeters, S.K.: The use of the normalized difference water index (NDWI) in the delineation of open water features. Int. J. Remote Sens. **17**(7), 1425–1432 (1996)
12. Hanqiu, X.: Modification of normalised difference water index (NDWI) to enhance open water features in remotely sensed imagery. Int. J. Remote Sens. **27**(14), 3025–3033 (2006)
13. Feyisa, G.L., Meilby, H., Fensholt, R., Proud, S.R.: Automated water extraction index: a new technique for surface water mapping using Landsat imagery. Remote Sens. Environ. **140**, 23–35 (2014)
14. Zhang, H., Yang, Y., Shen, H.: Detection of curvilinear structure in images by a multi-centered hough forest method. IEEE Access **6**, 22684–22694 (2018)
15. Yang, K., Li, M., Liu, Y., Cheng, L., Huang, Q., Chen, Y.: River detection in remotely sensed imagery using gabor filtering and path opening. Remote Sens. **7**(7), 8779–8802 (2015)
16. Mukherjee, J., Gupta, P., Gautam, H., Chintalapati, R.: Detection of narrow river trails with the presence of highways from Landsat 8 Oli images. In: CVIP 2022, Nagpur, India, pp. 659–673. Springer (2023)
17. Yang, K., Li, M., Liu, Y., Cheng, L., Duan, Y., Zhou, M.: River delineation from remotely sensed imagery using a multi-scale classification approach. IEEE J. Sel. Top. Appl. Earth Observations Remote Sens. **7**(12), 4726–4737 (2014)
18. Marín, D., Aquino, A., Gegúndez-Arias, M.E., Bravo, J.M: A new supervised method for blood vessel segmentation in retinal images by using gray-level and moment invariants-based features. IEEE Trans. Medical Imaging **30**(1), 146–158 (2010)
19. De, J., et al.: A graph-theoretical approach for tracing filamentary structures in neuronal and retinal images. IEEE Trans. Med. Imaging **35**(1), 257–272 (2015)
20. Ouzounis, G.K., Wilkinson, M.H.F.: Mask-based second-generation connectivity and attribute filters. IEEE Trans. Pattern Anal. Mach. Intell. **29**(6), 990–1004 (2007)
21. Van de Weijer, J., Van Vliet, L.J., Verbeek, P.W., van Ginkel, R.: Curvature estimation in oriented patterns using curvilinear models applied to gradient vector fields. IEEE Trans. Pattern Anal. Mach. Intell. **23**(9), 1035–1042 (2001)

22. Merveille, O., Naegel, B., Talbot, H., Najman, L., Passat, N.: 2D filtering of curvilinear structures by ranking the orientation responses of path operators (RORPO). Image Process. Line **7**, 246–261 (2017). https://doi.org/10.5201/ipol.2017.207
23. Strisciuglio, N., Azzopardi, G., Petkov, N.: Robust inhibition-augmented operator for delineation of curvilinear structures. IEEE Trans. Image Process. **28**(12), 5852–5866 (2019)
24. Chib, S., Greenberg, E.: Understanding the Metropolis-Hastings algorithm. Am. Stat. **49**(4), 327–335 (1995)
25. Byrd, R.H., Lu, P., Nocedal, J., Zhu, C.: A limited memory algorithm for bound constrained optimization. SIAM J. Sci. Comput. **16**(5), 1190–1208 (1995)
26. Lodhia, A., Sheffield, S., Sun, X., Watson, S.S: Fractional gaussian fields: a survey. Probab. Surv. **13**, 1–56 (2016)
27. Laville, B., Blanc-Féraud, L., Aubert, G.: Off-the-grid curve reconstruction through divergence regularization: an extreme point result. SIAM J. Imag. Sci. **16**(2), 867–885 (2023)

Peel and Pool: The Path to Mandala Perfection

Tusita Sarkar[1], Maitreyee Sar[2], and Partha Bhowmick[1](✉)

[1] Indian Institute of Technology, Kharagpur PIN 721302, India
pb@cse.iitkgp.ac.in

[2] Kalinga Institute of Industrial Technology, Bhubaneswar PIN 751024, India

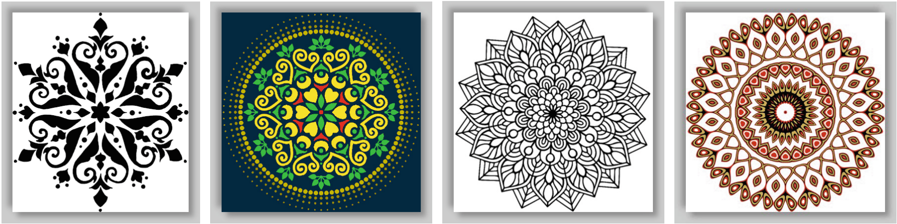
Abstract. Mandalas are renowned for their sacred symmetry, a principle that resonates deeply with human consciousness. This symmetry is manifested in the circular arrangement of their motifs. However, the circular layers derived from mandala images are often error-prone and unreliable for structural analysis. In contrast, using convex-hull layers and partitioning them into regular layers proves to be more reliable. We demonstrate this interesting fact through various illustrations and novel theoretical results. These findings, grounded in discrete geometry, facilitate the exploration of mandala structures using a novel peel-and-pool technique based on iterative convex-hull layers ('peeling') and regular layers ('pooling'). Our technique can rectify handcrafted or distorted mandalas, restoring their overall symmetry and enhancing their geometric harmony. These concepts offer fresh insights into the mathematical beauty of mandalas, with applications in art analysis, pattern recognition, and cultural studies. We validate the efficacy of our approach through experiments and visualizations.

Keywords: Mandala · Vedic art · Computational art · Art for society · Convex hull · Onion peeling · Symmetry

1 Introduction

The word 'mandala', derived from the ancient Sanskrit language, evocatively translates to 'circle' or 'discoid object'. In the rich spiritual tapestries of Hinduism and Buddhism, mandalas are celebrated as potent symbols of unity and cosmic interconnectedness. Far from being mere artistic creations, each mandala is a visual symphony of sacred symmetry and minimalist beauty. These intricate designs weave together interrelated motifs—circles, crescents, lotus petals, and other floral patterns—into harmonious, symmetrical arrangements that embody simplicity and balance. This meticulous symmetry invites profound introspection and spiritual connection, making mandalas powerful tools for meditation and reflection. The lasting appeal and sacred geometry of mandalas have crossed cultural boundaries, significantly influencing the art and culture of the Indian subcontinent and beyond [4,19,22].

A. Antonacopoulos et al. (Eds.): ICPR 2024, LNCS 15332, pp. 342–357, 2025.
https://doi.org/10.1007/978-3-031-78125-4_24



Class I Class I Class II Class II

Fig. 1. Two classes of mandalas. Our work is focused on Class I.

1.1 Origin and History of Mandalas

Vedic sacred symbols laid the foundation for mandalas [7,13]. Emerging from primitive art, mandalas evolved alongside written scripts as folk art expressions. Their decorative and auspicious nature makes them ideal for sacred platforms and wall paintings during worship and celebrations, captivating us with their enduring charm [17].

Mandalas, while largely maintaining their core structure, have evolved in response to socio-economic changes and have deeply ingrained in human cultural history. Hence, their aesthetic and socio-cultural significance is extensively documented in literary studies [4,7,13,15,17,22]. As heritage arts with universal appeal, they have influenced esteemed artists across the Indian subcontinent. Recent research by [7] traces the global evolution of their designs, particularly highlighting the journey of Bengal's traditional mandalas from folk art to symbols of cultural identity, as seen in places like Rabindranath Tagore's Santiniketan.

Mandalas are unique forms of art, deeply rooted in psychology. They evolve systematically from existing structures, reflecting psychic growth. As symbols of the 'Self' archetype, mandalas have shown great promise in psychotherapy, a potential first revealed over half a century ago [10]. Studies exploring the connection between art and the psyche have been ongoing since that time [2,3,9–11,18]. Recent research highlights the potential of mandalas in treating severe medical conditions such as Alzheimer's disease and cancer [1,6,12,14].

Constructed from a limited set of motifs, mandalas provide consistent designs that offer satisfaction through new combinations. In today's digital world, people globally are developing various computer-based methods to create mandalas. These methods often use loop generation and array grammars to facilitate mandala creation [5,16]. Recent advancements in this field are discussed in works such as [20,21].

1.2 Our Contribution

Our research is focused on analysis and correction of handcrafted mandalas. The digital images obtained from them lack symmetry or coherence due to manual

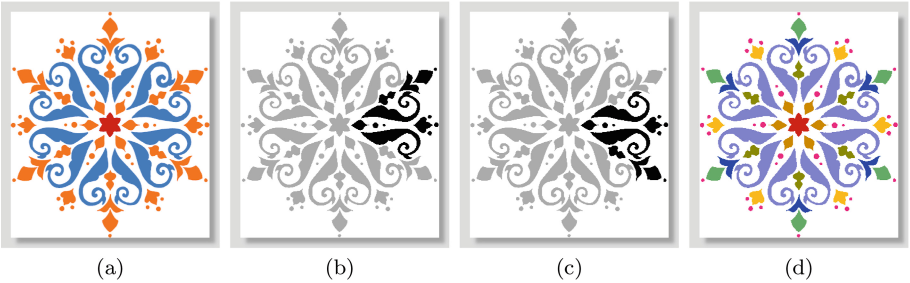


Fig. 2. (a) Central motif (red), symmetric motifs (saffron), and asymmetric motifs (blue).(b, c) Two possible ways of defining a sector (black). (d) Groups of motifs. (Color figure online)

errors during their creation or during digital acquisition. We deal with mandalas where the constituent elements are connected components distributed across symmetric sectors, which we refer to as Class I. There is another class in contrast, referred to as Class II, in which mandalas are essentially a curvilinear partition of a circular region. Some examples are given in Fig. 1. Henceforth, the term 'mandala' solely refers to Class I mandalas.

It may be noted that the existing works primarily focus on Class II mandala generation based on predefined models and do not address the analysis or refinement of handcrafted mandalas to achieve optimal configurations. In contrast, our research is centered on analyzing and streamlining handcrafted mandalas, which are manually drawn on paper or by a computer or graphics tablet. The novel contributions of our research can be summarized as follows.

1. Theoretically, we have introduced important facts and theorems necessary for the scientific analysis of mandalas.
2. We demonstrate how these theorems can be adapted to address practical challenges when working with digital images of handcrafted mandalas.
3. We have shown how an existing concept from computational geometry, known as 'onion peeling', can be transformed into a novel 'peel-and-pool' technique to extract motif arrangements and determine their symmetry.
4. This peel-and-pool technique is also used to correct structural errors in the digital versions of mandalas, ensuring precise and error-free representations.

2 Principles of Mandala Composition

Using the binary image of a mandala $\mathcal{M}$, its connected components are classified into distinct shapes known as *motifs* or *primitives*. Understanding these motifs is crucial for interpreting a mandala's composition. Refer to Fig. 2 for terminologies.

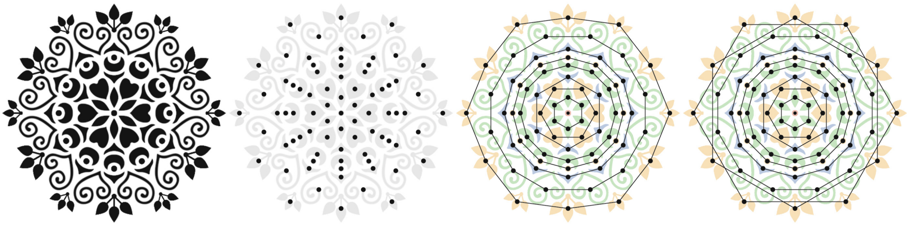

(a) Mandala image (b) Motif centroids (c) Hull layers (d) Regular layers

Fig. 3. Geometric characterization of a mandala (`mta602b1`) using motif centroids. The order of symmetry is 6, because for some regular layers, the number of centroids is 6, while for the others it is a multiple of 6.

At the mandala's center is the *central motif*, which appears exactly once in the entire mandala.[1] A *sector* is defined as the smallest set of connected components (excluding the central motif) whose rotation generates the entire mandala. The number of sectors determines its angle of rotational symmetry, or equivalently, the number of axes of reflection symmetry, referred to as the mandala's *order of symmetry* and denoted by $\psi_{\mathcal{M}}$.

In Fig. 2, the mandala consists of 79 connected components. Excluding the central motif, they can be grouped into 6 identical sectors, giving $\psi_{\mathcal{M}} = 6$. Each sector thus comprises 13 connected components, featuring 7 distinct motifs. A sector can be defined in more than one way, as shown in Fig. 2. In this example, only one motif is asymmetric and appears in a pair with reflection; we call it a *non-singular motif*. The 'disc' appears six times in different sizes, making it also non-singular. Every other motif occurs exactly once and is referred to as a *singular motif*. A non-singular motif may occur in different orientations and/or sizes.

2.1 Important Facts

The literature on mandalas, as referenced in §1, reveals key aspects of their composition, detailed in the following facts. It is crucial to note that these observations pertain strictly to mandalas of flawless precision. Handcrafted mandalas, and their digital counterparts, often exhibit noise and deviations, necessitating adjustments for accurate analysis and scientific interpretation. This is addressed later in §3. We first explicate here certain important facts related to mandala composition, with a reference to Fig. 3.

Fact 1. A mandala typically features a concentric, rotationally symmetric central motif, which acts as its focal point.

Fact 2. The focal point of a mandala being invariably its center, every symmetric-and-singular motif aligns its axis towards this central point.

[1] A mandala typically has a unique central motif, but occasionally it may have more than one, such as a disc encircled by a ring.

Fact 3. To maximize symmetry and harmony, mandalas typically feature a predominance of symmetric motifs over asymmetric ones.

Fact 4. An asymmetric motif typically appears in a non-singular manner, alongside its mirror image.Furthermore, when it is large in terms of area or perimeter, the mirroring occurs relative to the axis of symmetry of the corresponding sector.

Fact 5. Excluding the central motif, every motif occurs a multiple of $\psi_{\mathcal{M}}$ times throughout the mandala.Furthermore, if a motif $\mathcal{X}$ appears k times in sector j, labeled $\mathcal{X}_i^{(j)}$ for $i \in [1, k]$ and $j \in [1, \psi_{\mathcal{M}}]$, then, for every $i \in [1, k]$, the centroids of $\left\{\mathcal{X}_i^{(j)} : j \in [1, \psi_{\mathcal{M}}]\right\}$ form a regular $\psi_{\mathcal{M}}$-vertex convex polygon concentric with the center of the mandala.

The above facts collectively describe the composition of a mandala, ensuring its overall symmetry, sector symmetry, and motif symmetry. This symmetry converges towards the focal point, reflecting a mandala's symbolism and enhancing its spiritual significance for deep contemplation. These facts, along with additional geometric characterizations of motif centroids presented in the forthcoming section, are used to analyze the structural properties of a mandala, as discussed in §3.

2.2 Geometric and GCD-Based Characterization

We first introduce a few terminologies and their definitions needed for our work, with reference to Fig. 4. Consider a perfectly composed mandala $\mathcal{M}$, with its set of motif centroids (real points) denoted by S. A point p in S is said to be *covered* by a curve (e.g., a circle or a polygon) if p lies on that curve. A set $\mathcal{K}$ of curves is said to cover S if every point in S is covered by some curve in $\mathcal{K}$.

A polygon is *convex* if all its interior angles are less than 180°. For example, in Fig. 4(c), the topmost point is covered by the blue polygon, which is convex. A convex polygon is termed *regular* if all its sides are of equal length and all interior angles are also equal. In Fig. 4(c), all but the blue polygon are regular.

The *convex hull* of S is the smallest convex polygon that contains all the points in S. *Hull peeling* (a.k.a. 'onion peeling') is the technique of finding the convex hull of S, removing its vertices, and then repeatedly finding the convex hull of the remaining points until no points remain. In a similar manner, *circular peeling* is the technique of finding the smallest enclosing circle of S, and then repeatedly finding the next, until no points remain. Each circle found by this is termed a *circular layer*. The point set S in Fig. 4 has three hull layers and four circular layers. With the above terminologies in place, we herein introduce the following definitions. Let S denote the set of motif centroids of a perfect mandala excluding its central motif, and let $V(\cdot)$ denote the set of vertices of a polygon or of a collection of polygons, as applicable.

Definition 1. *Each convex hull found by the hull peeling of S is termed a* ***hull layer****, while each circle obtained by its circular peeling is a* ***circular layer****.*

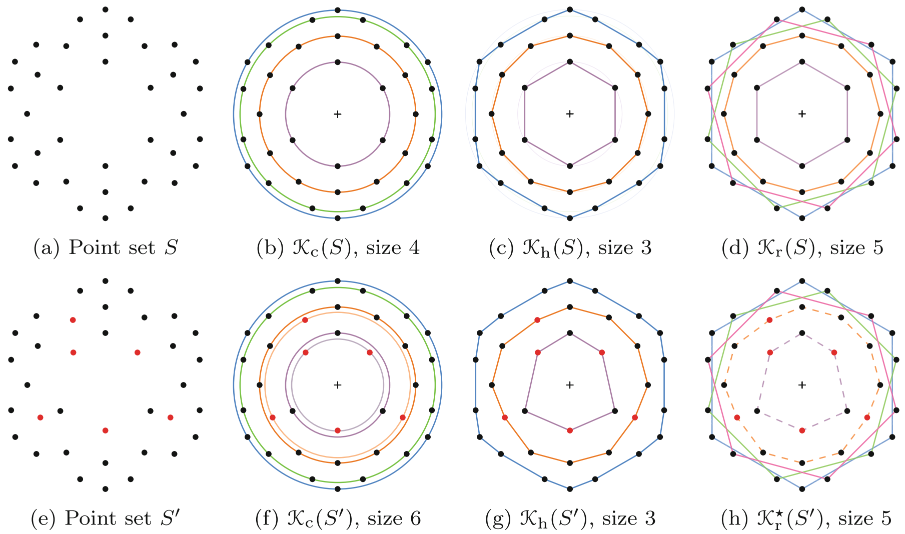

(a) Point set S
(b) $\mathcal{K}_{\mathrm{c}}(S)$, size 4
(c) $\mathcal{K}_{\mathrm{h}}(S)$, size 3
(d) $\mathcal{K}_{\mathrm{r}}(S)$, size 5
(e) Point set S'
(f) $\mathcal{K}_{\mathrm{c}}(S')$, size 6
(g) $\mathcal{K}_{\mathrm{h}}(S')$, size 3
(h) $\mathcal{K}_{\mathrm{r}}^{\star}(S')$, size 5

Fig. 4. Top row: A point set S and its three classes of cover, demonstrating Theorem 1 and the many-to-one maps from $\mathcal{K}_{\mathrm{r}}(S)$ to $\mathcal{K}_{\mathrm{h}}(S)$ and from $\mathcal{K}_{\mathrm{r}}(S)$ to $\mathcal{K}_{\mathrm{c}}(S)$. **Bottom row:** Corresponding covers for a point set S', which is obtained by perturbing a few points of the set S in (a); perturbed points are colored red; $\mathcal{K}_{\mathrm{r}}^{\star}(S')$ = pseudo-regular cover of S'.

Definition 2. *The* ***hull cover*** *of S, denoted by $\mathcal{K}_{\mathrm{h}}(S)$, is defined as the collection of its hull layers, while its* ***circular cover****, denoted by $\mathcal{K}_{\mathrm{c}}(S)$, is defined as the collection of its circular layers.*

Observe that a hull layer may not be regular, and in that case, its vertex set may be partitioned into a collection of fewest subsets so that each subset gives a regular polygon. For example, the blue polygon in Fig. 4(c) is not regular but its vertex set can be partitioned into three subsets, each forming a regular hexagon, as shown in Fig. 4(d). This minimum partition corresponds to a regular cover, as defined below.

Definition 3. *The* ***regular cover*** *of S, denoted by $\mathcal{K}_{\mathrm{r}}(S)$, is the smallest collection of regular polygons whose vertex sets are obtained by a partitioning of the vertex sets of the individual layers in $\mathcal{K}_{\mathrm{h}}(S)$. A polygon in $\mathcal{K}_{\mathrm{r}}(S)$ is termed a* ***regular layer*** *of S.*

In practical scenarios, a small perturbation of the points will increase the number of circular layers, with minimal or no effect on the number of hull layers. This is illustrated in Fig. 4(e–h). Notice that some of the polygons in Fig. 4(h) are not regular, and we refer to them as 'pseudo-regular' because they can be made regular with a little adjustment of the vertices. We revisit such scenarios later in §3.

We complete this section with the following theoretical findings, which are essential for analyzing the structural composition of mandalas. In the context that follows, "points in a layer" refers to the points of S covered by the layer. Similarly, "a layer covers another layer" means the points of S in the latter are covered by the former.

Theorem 1. *Each layer in $\mathcal{K}_\mathrm{h}(S)$ covers at least one layer in $\mathcal{K}_\mathrm{c}(S)$. Additionally, it may cover all or some points of other layers in $\mathcal{K}_\mathrm{c}(S)$.*

Proof. We provide a constructive argument. Let $C_1, \ldots, C_m$ be the layers in the collection $\mathcal{K}_\mathrm{c}(S)$, and $H_1, \ldots, H_n$ be those in $\mathcal{K}_\mathrm{h}(S)$, arranged in decreasing order of size in each collection. Observe that the polygon obtained by joining any set of points lying on any circle (in clockwise or counterclockwise order) is always convex. Hence, all points in the layer C_1 are covered by H_1. Moreover, H_1 may cover some points lying in another layer C_i, where $i \geq 2$. Proceeding in this manner, assume that $C_1, \ldots, C_i$ are all covered by $H_1, \ldots, H_i$, where $i \geq 1$. For the next circular layer C_{i+1}, there are three possibilities as follows:

i) C_{i+1} is covered by H_i alone.
ii) No point of C_{i+1} is covered by H_i.
iii) Some but not all points of C_{i+1} are covered by H_i.

For the first case, no new hull layer is needed. For the other two cases, we need a new hull layer, namely H_{i+1}, to cover C_{i+1}. □

Definition 3 implies that each layer P in $\mathcal{K}_\mathrm{r}(S)$ is covered by a unique layer in $\mathcal{K}_\mathrm{h}(S)$. Further, being a regular polygon, P is covered by a unique layer in $\mathcal{K}_\mathrm{c}(S)$. Combining these two facts, we get the following corollary.

Corollary 1. *There exists a many-to-one map from $\mathcal{K}_\mathrm{r}(S)$ to $\mathcal{K}_\mathrm{h}(S)$, and another from $\mathcal{K}_\mathrm{r}(S)$ to $\mathcal{K}_\mathrm{c}(S)$.*

Using the above result, we obtain the following two theorems that are important in the context of our work.

Theorem 2. *Let H be any layer in $\mathcal{K}_\mathrm{h}(S)$. Let $\mathcal{P} := \{P_i : i \in [1, m]\}$ be the collection of regular layers obtained from H. Let S_j be the subset of $V(\mathcal{P})$ covered by a circle C_j in $\mathcal{K}_\mathrm{c}(V(\mathcal{P}))$. Then, $\gcd\{|V(P_i)| : i \in [1, m]\}$ divides $|S_j|$.*

Proof. Since every layer P_i in $\mathcal{P}$ is a regular polygon, $V(P_i)$ gets covered by a unique circle in $\mathcal{K}_\mathrm{c}(V(\mathcal{P}))$. Let that circle be C_j. Additionally, due to the many-to-one map from $\mathcal{K}_\mathrm{r}(S)$ to $\mathcal{K}_\mathrm{c}(S)$ (Corollary 1), C_j may cover the vertices of some more polygons in $\mathcal{P}$. Denote by $\mathcal{P}_j := \{P_j^i : i \in [1, m_j]\}$ the layers in $\mathcal{P}$ covered by C_j. Let g_j denote $\gcd\{|V(P_j^i)| : i \in [1, m_j]\}$. Since the polygons in $\mathcal{P}$ are pairwise vertex-disjoint, and the sum of two or more numbers is divisible by their GCD, it follows that g_j divides $\sum\limits_{i=1}^{m_j} |V(P_j^i)| := |S_j|$.

Let n be the number of layers in $\mathcal{K}_\mathrm{c}(V(\mathcal{P}))$. Let g denote $\gcd\{g_j : j \in [1, n]\}$, which is identical with $\gcd\{|V(P_i)| : i \in [1, m]\}$. Since g divides g_j and g_j divides $|S_j|$ for $j \in [1, n]$, the result follows. □

The following theorem is an extension of the above theorem and states how the same GCD can be used to for a characterization of $\mathcal{K}_{\mathrm{h}}$. The notations H and $\mathcal{P}$ remain the same as above.

Theorem 3. *For any H in $\mathcal{K}_{\mathrm{h}}(S)$, $\gcd\{|V(P_i)| : i \in [1, m]\}$ divides $|V(H)|$.*

Proof. Let $g := \gcd\{|V(P_i)| : i \in [1, m]\}$. Observe that g divides $|V(P_i)|$ for $i \in [1, m]$, and hence divides $|V(\mathcal{P})| = |V(H)|$, as explained in the proof of Theorem 2. □

The above theorems are used to characterize the order of symmetry $\psi_{\mathcal{M}}$ for a perfect mandala $\mathcal{M}$, stated in the following theorem.

Theorem 4. *The number of points in S covered by every layer, whether in $\mathcal{K}_{\mathrm{c}}(S)$, $\mathcal{K}_{\mathrm{h}}(S)$, or $\mathcal{K}_{\mathrm{r}}(S)$, is divisible by $\psi_{\mathcal{M}}$.*

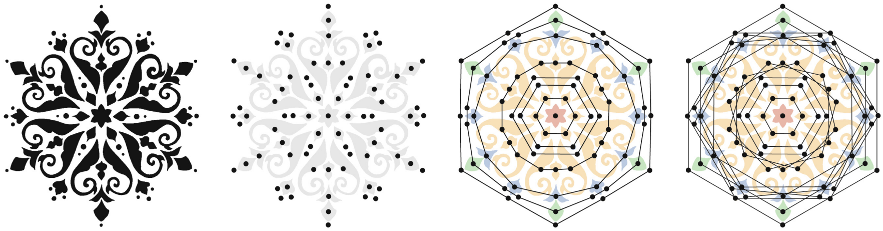
Proof. Consider any motif $\mathcal{X}$ in $\mathcal{M}$. Suppose that it appears k times in the j-th sector, denoted as $\mathcal{X}_i^{(j)}$, where $k \geqslant 1$, $i \in [1, k]$, and $j \in [1, \psi_{\mathcal{M}}]$. By Fact 5, we know that for every $i \in [1, k]$, the centroids of $\left\{\mathcal{X}_i^{(j)} : j \in [1, \psi_{\mathcal{M}}]\right\}$ form a regular $\psi_{\mathcal{M}}$-vertex convex polygon $P_i^{(j)}$ whose center coincides with that of the mandala. Clearly, for the set $P_{\mathcal{X}} := \left\{P_i^{(j)} : i \in [1, k]\right\}$, the corresponding vertices will be covered by a single circle C in $\mathcal{K}_{\mathrm{c}}(S)$. The circle C additionally may cover all the centroids corresponding to another motif. By Theorem 2, the number of points covered by C will be divisible by $\psi_{\mathcal{M}}$. By Theorem 3, and with a similar argument, this will also be true for every layer in $\mathcal{K}_{\mathrm{h}}(S)$ or $\mathcal{K}_{\mathrm{r}}(S)$. □

Absolute precision or perfection is not found in handcrafted or computer-generated mandalas. Hence, Theorem 4 is not directly applicable for determining the value of $\psi_{\mathcal{M}}$ in practice. However, it can be adapted to align with our needs, as discussed next.

3 Peel and Pool: Analyzing Digitized Mandalas

As we mentioned earlier, the digital image of a mandala is usually not perfect because its motifs are not accurately arranged in different layers. Further, the different instances of a particular motif in a particular layer are not exactly same. As a result, to identify and rearrange the motifs, we have to do a rectification. For this, we verify different measures associated with the motifs. Two values of any measure are considered the same if their relative difference (with respect to their sum) is within a small error margin, ε. This error margin is the same as the one used in Algorithm 1 (Line 4). Based on our experiments and empirical observations, we have set the value of ε to 0.05. A rationale for this is explained in §4.

The measures are taken for both the motifs and the layers. For motifs, the measures include their attributes as well as polar radii. For each layer, the measure depends on the type of layer. For example, a set of motifs are considered to



(a) Input image (b) Motif centroids (c) Hull layers (d) Regular layers

Fig. 5. Hull layers and regular layers in a mandala (`hda14a`). The order of symmetry is 6, as the number of centroids in each regular layer is either 6 or a multiple of 6.

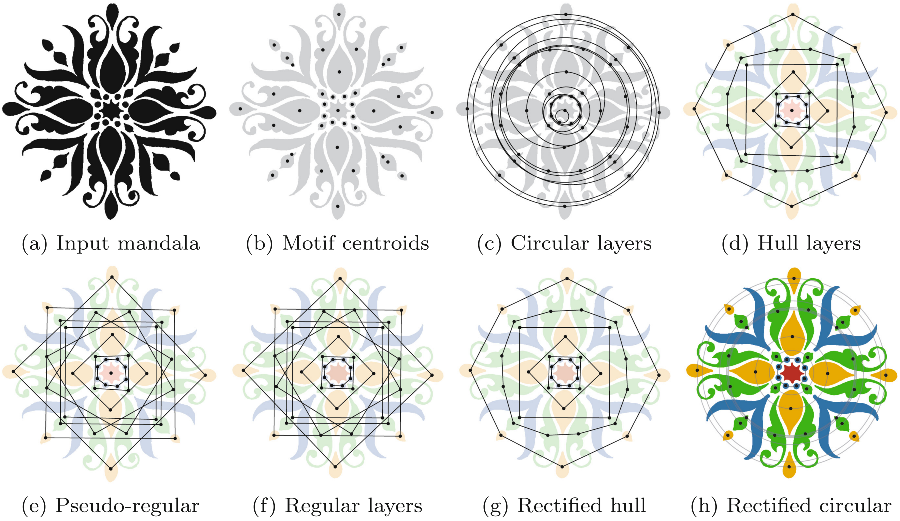
```
Algorithm 1: PeelAndPool($\mathcal{M}$)
1  Extract the hull layers $\{H_i : 1 \leqslant i \leqslant h\}$ by onion peeling of $\mathcal{M}$
2  for every layer $H_i$ do
3      set $V_i \leftarrow$ sequence of vertices of $H_i$
4      Compute a minimum partition of $V_i$: $\pi(V_i) \leftarrow \{V_{i,1}, \ldots, V_{i,m}\}$ such that
         $\forall\, 1 \leqslant s \leqslant m,\ \forall\, f \in \{\alpha, \beta, \gamma, \phi\},\ \frac{\max\{f_u : u \in V_{i,s}\} - \min\{f_u : u \in V_{i,s}\}}{\max\{f_u : u \in V_{i,s}\} + \min\{f_u : u \in V_{i,s}\}} \leqslant \varepsilon$
5      for $j \in [1, m]$ do
6          Compute a minimum partition of $V_{i,j}$: $\pi(V_{i,j}) \leftarrow \{V_{i,j,1}, \ldots, V_{i,j,t}\}$
             such that each $V_{i,j,k}$ is pseudo-regular and the twin motifs are in
             different parts // pseudo-regular layers
7          for each $V \in \pi(V_{i,j})$ do
8              $r \leftarrow \frac{1}{|V|} \sum_{u \in V} r_u$ // $r_u$ = polar radius of $u$
9              Fix the motif centroids uniformly on a circle of radius $r$
                 // regular layers
10     $g_i \leftarrow \gcd\{|V| : (V \in \pi(V_{i,j})) \wedge (1 \leqslant j \leqslant m)\}$ // Theorem 2
11 $\psi_{\mathcal{M}} \leftarrow \gcd\{|g_i| : 1 \leqslant i \leqslant h\}$ // Theorem 4
```

be in the same circular layer if the polar radii of their centroids, taken pairwise, have a relative difference of at most ε. For the hull layers, the interior angles of the polygons are used as measures, and for the regular layers, the edge lengths are considered to determine whether motifs belong to the same layer. Figure 5 shows an example, illustrating hull layers and regular layers.

3.1 Identification of Layers and Order of Symmetry

To identify hull layers and regular layers, and to determine the order of symmetry, we use the theoretical concepts discussed in §2.2. The main steps are given in Algorithm 1. After computing the motif centroids, the hull layers are extracted, as shown in Line 1. Then the following features of the motifs are extracted for each hull layer to partition it into pseudo-regular layers:

1. α = area of a motif, measured by its number of pixels.


(a) Input mandala (b) Motif centroids (c) Circular layers (d) Hull layers

(e) Pseudo-regular (f) Regular layers (g) Rectified hull (h) Rectified circular

Fig. 6. Demonstration of our algorithm on `hdalpona9`. The mandala is imperfect because the collection of circular layers changes after rectification. All its layers are rectified by our algorithm to obtain a structurally perfect mandala.

2. $\beta =$ *compactness ratio*, defined as the ratio of the square of the perimeter to the area of a motif [8].
3. $\gamma =$ *solidity*, defined as the ratio of the area of the motif to the area of its convex hull.
4. $\phi =$ interior angle of the hull layer (polygon) measured at the motif centroid.

The pseudo-regular layers are obtained in two stages, shown in Line 4 and Line 6. In Line 4, the above four features are used to partition a hull layer, while in Line 6 the partition is further subdivided based on edge lengths and twin motifs. We call two asymmetric motifs $\mathcal{X}'$ and $\mathcal{X}''$ a 'twin' if they are approximately identical under reflection, i.e., their relative difference (ratio of difference to $\alpha(\mathcal{X}')+\alpha(\mathcal{X}'')$) after reflecting one of them is at most ε. To compare the edge lengths we also use ε as a tolerance. We ensure that the maximum normalized standard deviation (ratio of standard deviation to the average) of all edge lengths in every pseudo-regular layer is at most ε. Using the number of motifs in these pseudo-regular layers, we determine the value of $\psi_{\mathcal{M}}$ in Line 11.

A demonstration is given in Fig. 6. After obtaining the hull layers, the pseudo-regular layers and the order of symmetry are derived.

3.2 Mandala Rectification

After gathering all the pseudo-regular layers of $\mathcal{M}$ in Line 6 of Algorithm 1, we proceed to correct the imperfections in $\mathcal{M}$. As shown in Lines 7–9, we adjust the polar coordinates (r, θ) of the centroids for each pseudo-regular layer to equalize

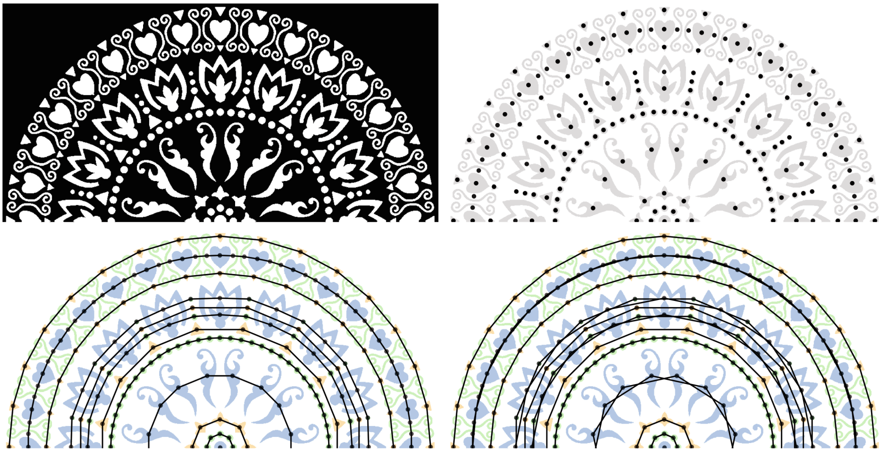


Fig. 7. Result obtained by our algorithm on the mandala `mta802a2` (one half shown). This is found to be almost perfect by our algorithm ($\sigma_{\max} = 0.008$, Table 1).

their edge lengths so that the layer is regular. This is done by setting the polar radii of the centroids to the same value given by the arithmetic mean $\bar{r}$ of all the polar radii within the layer. We then identify the centroid $u \in V$ whose polar radius r_u is closest to $\bar{r}$. By aligning u to its nearest line of symmetry and setting $r_u = \bar{r}$, we position the remaining centroids of V with an angular spacing of $\frac{360^\circ}{|V|}$ between every two consecutive centroids. This process is demonstrated in Fig. 6.

4 Experiments and Results

The algorithm is completely implemented in Python. The code is executed in a Lenovo IdeaPad Flex 5 laptop having 11th Gen Intel Core i7-1165G7 CPU@2.80GHz and 16GB RAM. The mandalas in our test dataset are either handcrafted by professional artists on paper or on a tablet using a stylus, or they are computer-generated. The physical sizes of the mandalas range from 4 inches to 12 inches in diameter. The images of the ones drawn on paper are acquired by cameras in ordinary mobile phones. Visual results on some of them are presented in Figs. 6, 7, 8 and 9.

In Table 1, we have presented the results for some images to assess the performance of our algorithm. It contains the image details including geometric properties of the mandalas and CPU times for different stages of our algorithm. An important property of a mandala image is the amount of perfection in its structural composition captured in its regular layers. For every pseudo-regular layer, we use the normalized standard deviation of its edge lengths, measured by the usual standard deviation divided by their average. We consider $\sigma_{\min}$, $\sigma_{\max}$, and σ_{avg} as the respective minimum, maximum, and average values of the normalized standard deviations over all pseudo-regular layers of the mandala.

Table 1. Statistical details demonstrating the performance of the algorithm on different mandalas. $T_{\text{prim}}, T_{\text{hull}}, T_{\text{regu}}$ = respective CPU times (in seconds) for extraction of motifs and their features, for extracting hull layers and computing $\psi_{\mathcal{M}}$, and for pooling pseudo-regular layers. T_{rect} = time for rectification. $T = T_{\text{prim}} + T_{\text{hull}} + T_{\text{regu}} + T_{\text{rect}}$.

Image	name	n	$\psi_{\mathcal{M}}$	size	$\lvert\mathcal{K}_{\text{h}}(S)\rvert$	$\lvert\mathcal{K}_{\text{r}}(S)\rvert$	$\sigma_{\min}$	$\sigma_{\max}$	σ_{avg}	T_{prim}	T_{hull}	T_{regu}	T_{rect}	T
1	hda9	41	4	1467 × 1500	6	9	0.006	0.027	0.015	0.024	0.010	0.002	10.645	10.681
2	mta602b1	79	6	536 × 536	8	10	0.001	0.0035	0.002	0.020	0.018	0.003	1.408	1.449
3	hda14	79	6	1000 × 1000	7	13	0.001	0.009	0.004	0.018	0.016	0.003	3.324	3.361
4	mta803c	193	8	1000 × 1000	10	13	0.002	0.011	0.006	0.017	0.046	0.005	3.251	3.319
5	mta12d1	217	8	1000 × 1000	8	8	0.01	0.022	0.014	0.019	0.052	0.005	3.237	3.313
6	mta802a2	353	8	1000 × 1000	11	15	0	0.008	0.003	0.021	0.082	0.008	3.411	3.522
7	mta607c2	529	6	1000 × 1000	14	16	0.001	0.013	0.004	0.023	0.154	0.012	3.31	3.499
8	mta16j1	889	8	594 × 594	19	22	0	0.012	0.006	0.027	0.295	0.022	3.225	3.569

From our experiments we can see that the value of $\sigma_{\min}$ is quite low for all the mandala images, whether they are handcrafted or computer-generated. On the contrary, the value of $\sigma_{\max}$ differs from one mandala to the other quite significantly depending on the level of accuracy. In computer-generated instances, the accuracy is more, whereas it is usually not so for handcrafted ones. For example, for the first image in Table 1, the value of $\sigma_{\max}$ is much higher than that in the second image because the former is handcrafted by an artist but the latter is computer-generated. Observe that for all the mandalas in this table, the value of $\sigma_{\max}$ is within 0.05, which provides a justification of setting the value of ε to 0.05, as mentioned earlier in §3. In Fig. 7, we can see an example of an almost perfect mandala, which is evident from its low value of $\sigma_{\max}$ in Table 1.

Regarding the execution time of our algorithm, an important observation is that the time to extract motifs and their features and the time to rectify distortions depend on the size of the image. This is evident from the fact that hdalpona9 requires significantly longer time to be analyzed and corrected, in spite of having fewer motifs. Additionally, the time to rectify is influenced by the degree of distortion present in the mandala, as seen particularly with the fourth image, which has a high $\sigma_{\max}$.

Observe that most of the execution time is consumed by the rectification process. This step is lengthy because it evaluates and corrects each pixel in every motif of the mandala. In contrast, the time required for extracting hull and regular layers is much shorter, as it only involves the motif centroids. The CPU execution time for layer extraction primarily depends on the number of motifs.

In Fig. 9, we provide a step-by-step demonstration of our algorithm on a larger and denser computer-generated mandala. The mandala is distorted to test our algorithm's performance. The improvement is evident from the decreased number and better alignment of the circular layers after rectification, as shown in Fig. 9(c, g).

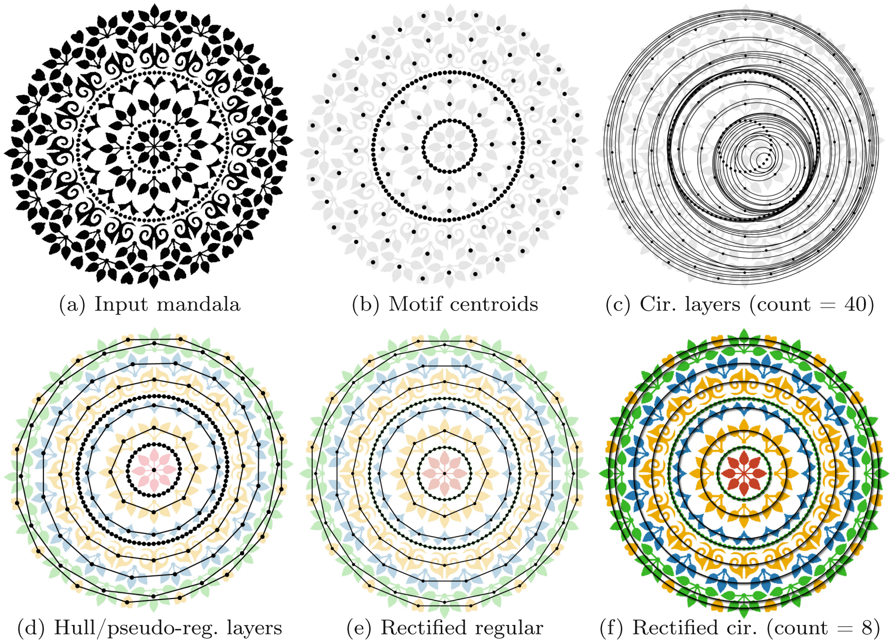
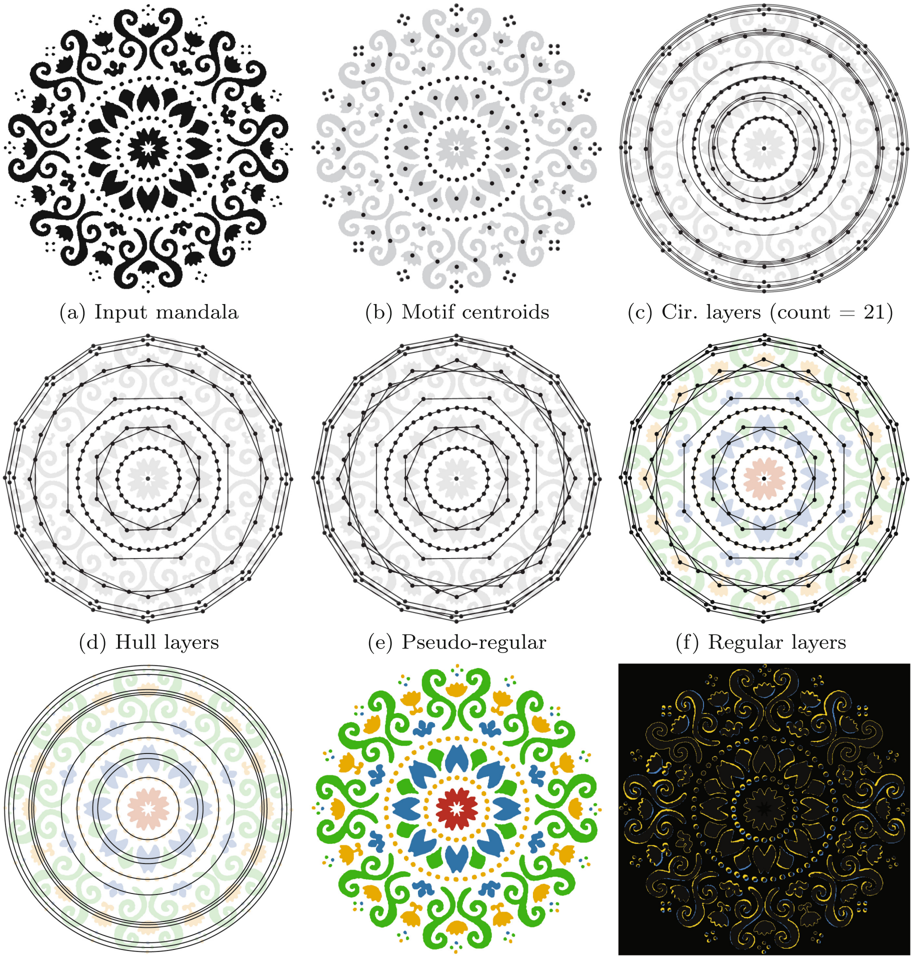

(a) Input mandala
(b) Motif centroids
(c) Cir. layers (count = 40)
(d) Hull/pseudo-reg. layers
(e) Rectified regular
(f) Rectified cir. (count = 8)

Fig. 8. Results by our algorithm on `mta12d1`. It accurately identifies the hull layers in a mandala, even in the presence of distortions. Circular layers are grossly incorrect in the original image but corrected in the final result.

The correctness of our algorithm depends on whether it can extract the hull layers and the pseudo-regular layers correctly. In case of excessive distortions and noise, these layers may not be correctly identified, which is illustrated through an example in Fig. 10. From this example, it is apparent that the coalescence of certain motifs causes our algorithm to fail in correctly extracting the hulls.

5 Concluding Notes

To the best of our knowledge, no existing work analyzes and characterizes the structural composition of mandalas, particularly when they are imperfect. The novel technique introduced in this paper is based on the classical concept of repeated convex hulls for peeling a point set, offering a high-level description of a mandala that can be applied in subsequent computational art applications. This technique can be further explored to address mandalas with abnormal noise and distortions and may be enhanced by incorporating suitable machine learning techniques. Additionally, the contours of the motifs, often quite jagged, may be smoothed in the final mandala after rectification. Furthermore, the technique's applicability can be extended to other similar art forms such as *Alpona*, *Rangoli*, and *Kolam*.

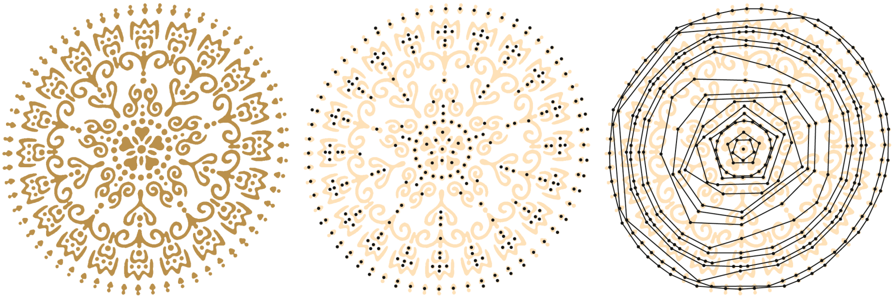


(a) Input mandala (b) Motif centroids (c) Cir. layers (count = 21)

(d) Hull layers (e) Pseudo-regular (f) Regular layers

(g) Rectified cir. (count = 11) (h) Rectified mandala (i) Difference

Fig. 9. Demonstration of our algorithm on `mta803c`. It is identified as an imperfect mandala and its pseudo-regular layers are corrected to regular layers to obtain a rectified mandala. (i) Yellow = pixels in the motifs of the rectified mandala but not in the original one; blue = those in the original mandala but not in the rectified one.



(a) Input mandala (b) Motif centroids (c) Hull layers

Fig. 10. The mandala `mta503b` with excessive noise.In such images, our algorithm does not yield perfect results.The main reason is that some motifs coalesce together or break apart, resulting in incorrect centroids.

Acknowledgements. This work is partially supported by a research grant from the GKF project under the *Academy of Classical and Folk Arts*, IIT Kharagpur, India.We extend our gratitude to Dr. Mukund Padmanabhan for his ongoing financial support and continuous encouragement.We also thank Mr. Tushar Abhishek for providing a tool to create computer-generated mandalas.

References

1. Akbulak, F., Can, G.: Effectiveness of mandala coloring in reducing anxiety in women with early-stage breast cancer receiving chemotherapy for the first time. EXPLORE **19**(1), 42–47 (2023)
2. Allchin, W.: Mandalas and manipulators: a Jungian insight into order and chaos and its impact on the adolescent. J. Adolesc. **3**(1), 1–10 (1980)
3. Bush, C.A.: Dreams, mandalas, and music imagery: therapeutic uses in a case study. Arts Psychother. **15**(3), 219–225 (1988)
4. Chaitanya, K.: A History of Indian Painting. Abhinav Publications, India (1976)
5. David, N.G., Thamburaj, R., Thomas, D.G., Balamurugan, B., Samuel, S.: Graph grammars for Kolam patterns and honeycomb patterns. Int. J. Math. Sci. **6**, 355–367 (2007)
6. Elkis-Abuhoff, D., Gaydos, M., Goldblatt, R., Chen, M., Rose, S.: Mandala drawings as an assessment tool for women with breast cancer. Arts Psychother. **36**(4), 231–238 (2009)
7. Ghosh, S.: Design Movement in Tagore's Santiniketan: Alpana–An Experiment in Aestheticism. Niyogi Books, New Delhi (2019)
8. Gonzalez, R.C., Woods, R.E.: Digital Image Processing. Prentice Hall, India (2002)
9. Ireland, M.S., Brekke, J.: The mandala in group psychotherapy: personal identity and intimacy. Arts Psychother. **7**(3), 217–231 (1980)
10. Jung, C.G., Jaffé, A., Winston, R., Winston, C.: Memories, Dreams, Reflections. Pantheon Books, New York (1963)
11. Kainen, P.C.: Graphs, groups and mandalas. In: Hammer, P.L. (ed.) Combinatorics 79, Annals of Discrete Mathematics, vol. 9, p. 125. Elsevier (1980)

12. Kim, H., Kim, S., Choe, K., Kim, J.S.: Effects of mandala art therapy on subjective well-being, resilience, and hope in psychiatric inpatients. Arch. Psychiatr. Nurs. **32**(2), 167–173 (2018)
13. Mitra, A.: Bharater chitrakala (artworks of India). In: Encyclopaedia Britannica (Vol. 12) (1986)
14. Moharamkhani, M., Rassouli, M., Mojen, L.K., Respini, D., Aghebati, A., Ashrafizadeh, H.: Assessing effects of mandala painting on anxiety of 9–14-year-old children with cancer. Ad. Integr. Med. **10**(1), 8–14 (2023)
15. Murugan, I., Perumal, V., Kamarudin, K.: Challenges in the practice of traditional Kolam among Indian women in the Klang valley. Alam Cipta **14**, 58–68 (2021)
16. Pradella, M., Cherubini, A., Crespi Reghizzi, S.: A unifying approach to picture grammars. Inf. Comput. **209**(9), 1246–1267 (2011)
17. Sengupta, P.: Meyeli brater alpanar itihaas. In: Mitra, S.K. (ed.) Meyeli Brata Bishoye. Loksanskriti Gabeshana Parishad, Kolkata (1986)
18. Slegelis, M.H.: A study of Jung's mandala and its relationship to art psychotherapy. Arts Psychother. **14**(4), 301–311 (1987)
19. Tanabe, W.J.: Japanese mandalas: representations of sacred geography. Jpn. J. Religious Stud. **28**(1–2), 186–188 (2001)
20. Xu, S., Zhang, Y., Yan, S.: Automatic mandala pattern design and generation based on COOM framework. J. Comput. Lang. **72**, 101138 (2022)
21. Zhang, J., Zhang, K., Peng, R., Yu, J.: Parametric modeling and generation of mandala thangka patterns. J. Comput. Lang. **58**, 100968 (2020)
22. Zimmer, H.: Myths and Symbols in Indian Art and Civilization (Ed. Joseph Campbell) Princeton University Press, New Jersey (1972)

MCANet: Multimodal Caption Aware Training-Free Video Anomaly Detection via Large Language Model

Prabhu Prasad Dev, Raju Hazari, and Pranesh Das(✉)

Machine Learning Laboratory, Department of Computer Science and Engineering,
National Institute of Technology, Calicut, India
prasaddev97@gmail.com
{rajuhazari,praneshdas}@nitc.ac.in

Abstract. Towards Video Anomaly Detection (VAD), existing methods require labor-intensive data collection and model retraining, making them costly and domain-specific. The proposed method, termed as Multimodal Caption Aware Network (MCANet), introduces a novel paradigm that identifies anomalies in video sequences without requiring prior domain knowledge. This training-free VAD approach dynamically generates and analyzes textual descriptions of video frames by utilizing off-the-shelf vision-language model (VLM), audio-language model (ALM) and large language model (LLM). MCANet has four primary modules. The first module utilizes image-text similarities to clean noisy captions generated by the image captioning model, while the second module applies audio-text similarities to refine noisy captions produced by the audio captioning model. The third module employs a LLM to consolidate scene dynamics over time. Finally, the fourth module enhances the results by aggregating scores from semantically similar frames based on video-text similarity. To validate the effectiveness of the proposed method, experiments are conducted on two large-scale benchmark datasets (UCF-Crime and XD-Violence). Experimental results demonstrate that MCANet surpasses existing unsupervised and one-class approaches without requiring any training or data collection.

Keywords: Video Anomaly Detection · Large Language Model · Vision Language Model · Audio Language Model · Multimodal captions

1 Introduction

Video Anomaly Detection (VAD) has emerged as a pivotal method in identifying abnormal events within video sequences, garnering significant attention due to its broad applications in public safety and video content analysis. VAD methodologies are broadly classified based on the annotation type of the training data into unsupervised, weakly-supervised, and fully-supervised categories. Unsupervised approaches are designed to train on normal or unlabeled videos, while

A. Antonacopoulos et al. (Eds.): ICPR 2024, LNCS 15332, pp. 358–375, 2025.
https://doi.org/10.1007/978-3-031-78125-4_25

weakly supervised methods utilize video-level labels for both normal and abnormal videos. Though less prevalent due to the intensive requirement for frame-by-frame annotations, fully-supervised methods offer precise anomaly detection. Recent advancements in VAD leverage multi-modal large language models (MLLMs) pre-trained on extensive datasets, significantly enhancing detection accuracy and efficiency. The practical applications of VAD span various domains, including intelligent manufacturing, traffic surveillance, and public security. Conventional VAD techniques predict anomaly scores for each frame in the video sequence, with higher scores indicating a greater likelihood of anomalies. These methodologies emphasize automation, progressively advancing towards more robust and reliable VAD systems.

Existing state-of-the-art(SOTA) methods necessitate a training procedure to establish an accurate VAD system, which entails certain limitations. A primary concern is generalization: a VAD model trained on a specific dataset tends to underperform in videos recorded under different conditions (e.g., daylight versus night scenes). Moreover, data collection poses a significant challenge, particularly in application domains like video surveillance, where privacy issues can impede data acquisition. Therefore, to address these issues, a novel training-free approach is introduced that identifies anomalies in video sequences without requiring prior domain knowledge. The comaparison of the proposed method with existing methods is showcased in Fig. 1.

This paper aims to address these challenges by proposing the training-free **M**ulti-modal **C**aption **A**ware **Net**work (**MCANet**), which leverages pre-trained vision-language models (VLMs) and large language models (LLMs) for VAD. MCANet employs an off-the-shelf captioning model to generate textual descriptions for each video frame and integrates an audio-text captioning model to capture audio cues that contribute to the detection process. The combination of visual-text and audio-text captions provides a richer context, enhancing the model's ability to identify anomalies. To address noise in the captions, a cleaning process based on cross-modal similarity between captions and frames is introduced. Additionally, to capture scene dynamics, an LLM summarizes captions within a temporal window, generating an anomaly score for each frame, which is refined by aggregating anomaly scores from frames with similar temporal summaries. MCANet is evaluated on two large-scale benchmark datasets UCF-Crime and XD-Violence, demonstrating that this training-free approach outperforms unsupervised and one-class VAD methods, showing that VAD can be effectively addressed with no training and no data collection.

In summary, the contributions of this paper are as four-fold:

1. A novel training-free VAD approach is introduced that identifies anomalies in video sequences without requiring prior domain knowledge.
2. A Multi-modal Caption Aware Network (MCANet) is introduced as the first language-based method for training-free VAD, utilizing Large Language Models (LLMs) to detect anomalies solely from scene descriptions.

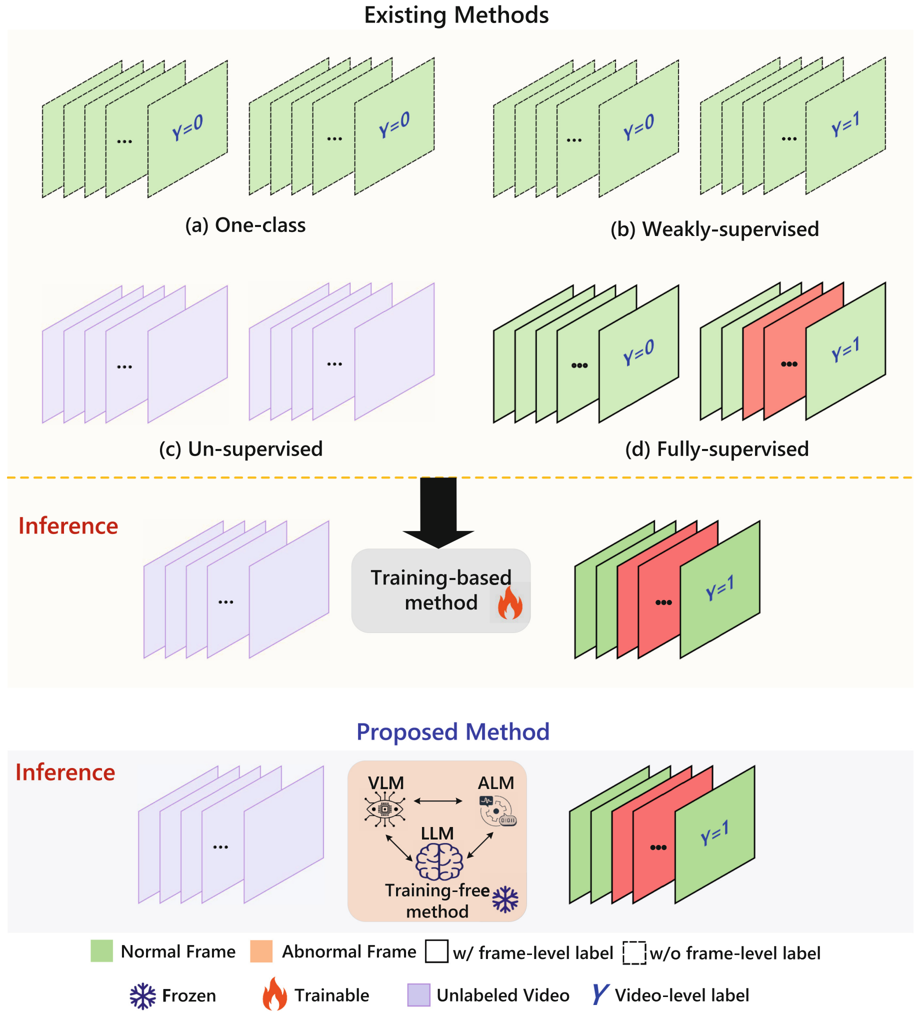


Fig. 1. Comparison with existing methods

3. MCANet consists of four main modules: the first two clean and refine noisy captions using image-text and audio-text similarities, respectively. The third module consolidates scene dynamics using an LLM, while the fourth aggregates scores from semantically similar frames based on video-text similarity.
4. Experiments results demonstrate that MCANet achieves competitive results on UCF-Crime and XD-Violence compared to unsupervised and one-class VAD methods without task-specific supervision or training.

2 Related Work

2.1 Video Anomaly Detection

Unsupervised video anomaly detection aims to identify unusual events in video data without requiring labeled training samples. This approach is crucial in scenarios where acquiring annotated data is challenging or impractical. For instance, Zhao et al. [1] utilized LSTM to leverage the spatio-temporal correlations among consecutive frames in normal videos. Lee et al. [2] employed Vision Transformer with spatio-temporal contextual prediction streams to enhance performance on the VAD task. These methods typically rely on the assumption that anomalies are rare and distinct from normal patterns. These unsupervised methods [3–6] offers scalability and adaptability by learning from unlabeled data, it often falls short in accuracy and robustness due to the lack of explicit guidance on what constitutes anomalies. Consequently, Weakly supervised video anomaly detection has emerged as a highly attractive and widely adopted technique in the research community. This approach leverages limited labeled data, often only at the video level, to learn discriminative features that can distinguish between normal and abnormal events. For instance, Sultani et al. [7] pioneered the use of a weakly supervised multiple instance learning (MIL) framework for video anomaly detection, where videos are considered as bags of segments. Their model is designed to assign higher anomaly scores to anomalous segments and lower scores to normal ones. However, their approach does not account for the temporal relationships between video segments and struggles to extract features that effectively discriminate between normal and anomalous snippets. Addressing these limitations, Huang et al. [8] enhanced the MIL framework by introducing a discriminative feature encoder that improves the distinction between normal and anomalous segments, along with a temporal feature aggregator that captures long-term dependencies across video sequences. Ullaha et al. [9] also developed a novel anomaly detection based on deep convolutional neural network and multi-head sequential attention-based temporal mechanism. Recently, Karim et al. [10] proposed an online video anomaly detection model that utilizes an end-to-end methodology. This approach enables the automatic extraction of features directly from raw videos, contrasting with traditional methods that depend on separate feature encoders and classifiers in ad-hoc settings.

2.2 Video-Based Large Language Models (VLLMs)

Video-based Large Language Models (VLLMs) have shown substantial progress in understanding and reasoning over language and visual content, reflecting the growing interest in applying Large Language Models (LLMs) to multimodal challenges. These models integrate the dynamic visual information of videos with the rich contextual details provided by textual descriptions. The efficacy of VLLMs in tasks such as video captioning [11], video understanding [12,13], image patch summarization [14], and interactive learning [15] highlights their potential to transform how machines understand and interact with complex, real-world data. Bain et al. [16] proposed an end-to-end dual encoder architecture for text-video retrieval. Li et al. [17] introduced a chat-centric video dialogue system by leveraging LLM. Zhang et al. [18] developed an audio-visual language framework empowering LLM for effective video understanding. Chen et al. [13] utilized LLM to handle video understanding tasks seamlessly. Lin et al. [19] presented a robust large vision language model that integrates visual representations into the language feature space, enabling the model to effectively interpret both images and videos. He et al. [20] devised a memory bank to store historical video content, enabling effective long-term video analysis without exceeding LLMs' context length or GPU memory limits. This approach significantly enhanced performance in tasks like video question answering and captioning, surpassing state-of-the-art models.

3 Methodology

In this section, the Video Anomaly Detection (VAD) problem is formalized, and the proposed training-free approach is outlined. The capabilities of Large Language Models (LLMs) in assigning anomaly scores to video frames are then examined. The proposed method framework is presented in Fig. 2.

3.1 Problem Formulation

Given a test video $\mathbf{V} = [f_1, \ldots, f_N]$ of N frames, traditional VAD methods aim to learn a model g, which can classify each frame $f \in \mathbf{V}$ as either normal (score 0) or anomalous (score 1), i.e., $g : (\mathcal{I} \times \mathcal{A})^N \rightarrow [0, 1]^N$ with $\mathcal{I}$ being the image space and $\mathcal{A}$ being the audio space. g is usually trained on a dataset $\mathcal{D}$ that consists of tuples in the form $(\mathbf{V}, y)$. Depending on the supervision level, y can be either a binary vector with frame-level labels (fully-supervised), a binary video-level label (weakly-supervised), a default one (one-class), or absent (unsupervised). However, in practice, it can be costly to collect y as anomalies are rare, and $\mathbf{V}$ itself due to potential privacy concerns. Moreover, both label and video data may need regular updates due to evolving application contexts.

In contrast to traditional methods, this paper presents a novel approach to VAD, termed training-free VAD. In this innovative setup, the goal is to estimate the anomaly score for each $f \in \mathbf{V}$ using only pre-trained models during inference, eliminating the need for any training or fine-tuning involving a training dataset $\mathcal{D}$.

3.2 MCANet

A novel approach, Multi-modal Caption Aware-Video Network (**MCANet**) for video anomaly detection, is proposed by leveraging advancements in Large Language Models (LLMs). The framework is depicted in Fig. 2. Recognizing the emerging use of LLMs in VAD, the initial step involves evaluating their ability to generate anomaly scores based on textual descriptions of video frames. A

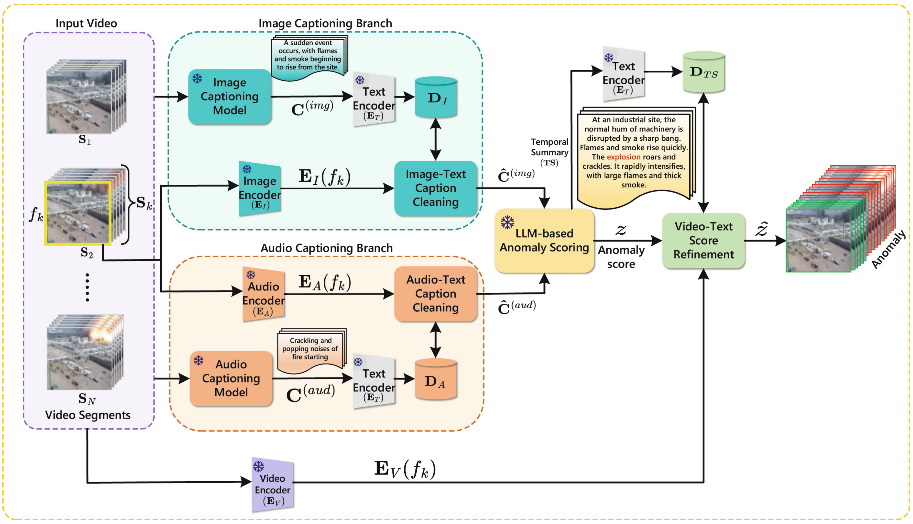

Input Video
$\mathbf{S}_1$
f_k
$\mathbf{S}_k$
$\mathbf{S}_2$
$\mathbf{S}_N$
Video Segments
Image Captioning Branch
Image Captioning Model
A sudden event occurs, with flames and smoke beginning to rise from the site.
$\mathbf{C}^{(img)}$
Text Encoder $(\mathbf{E}_T)$
$\mathbf{D}_I$
Image Encoder $(\mathbf{E}_I)$
$\mathbf{E}_I(f_k)$
Image-Text Caption Cleaning
$\hat{\mathbf{C}}^{(img)}$
Audio Captioning Branch
Audio Encoder $(\mathbf{E}_A)$
$\mathbf{E}_A(f_k)$
Audio-Text Caption Cleaning
$\hat{\mathbf{C}}^{(aud)}$
Audio Captioning Model
Crackling and popping noises of fire starting
$\mathbf{C}^{(aud)}$
Text Encoder $(\mathbf{E}_T)$
$\mathbf{D}_A$
Video Encoder $(\mathbf{E}_V)$
$\mathbf{E}_V(f_k)$
Temporal Summary (TS)
Text Encoder $(\mathbf{E}_T)$
$\mathbf{D}_{TS}$
At an industrial site, the normal hum of machinery is disrupted by a sharp bang. Flames and smoke rise quickly. The explosion roars and crackles. It rapidly intensifies, with large flames and thick smoke.
LLM-based Anomaly Scoring
z
Anomaly score
Video-Text Score Refinement
$\hat{z}$
Anomaly

Fig. 2. Proposed Framework

state-of-the-art image-to-text captioning model, θ_I, specifically FUSECAP [21], is utilized to convert each frame $f \in \mathbf{V}$ into a textual description, along with an audio-to-text captioning model, θ_A, specifically HTSAT, to convert each frame $f \in \mathbf{V}$ into a textual description. Anomaly score estimation is then addressed as a classification task, instructing an LLM, θ_{LLM}, to choose a score from a list of uniformly sampled values ranging from 0 to 1, with 0 indicating normal and 1 indicating an anomaly. The anomaly score is computed as follows:

$$\theta_{LLM}(P_C \circ P_O \circ \theta_I(f) \circ \theta_A(f)) \tag{1}$$

Here, P_C is a context prompt that provides priors to the LLM regarding VAD, and P_O guides the LLM on the desired output format for automated text parsing. The symbol $\circ$ denotes the text concatenation operation. P_C is used to simulate the perspective of a potential user of the VAD system, such as a security analyst, to enhance the LLM's effectiveness in generating the output. For instance, a P_C prompt could be: "If you are a security analyst, how would you rate the event described on a scale from 0 to 1, where 0 represents a normal event and 1 denotes an event with anomalous activities?" It is important to note that P_C does not inherently encode any specific information about the type of anomalies but rather provides context.

MCANet breaks down the VAD function f into seven components. As in the initial study, the first two are the captioning module θ_C and θ_A, which maps images to textual and audio to textual descriptions in the language space $\mathcal{T}$ respectively. Mathematically, $\theta_C : \mathcal{I} \rightarrow \mathcal{T}$ and $\theta_A : \mathcal{A} \rightarrow \mathcal{T}$ and the LLM θ_{LLM} which generates text from language queries, i.e., $\theta_{LLM} : \mathcal{T} \rightarrow \mathcal{T}$. The remaining elements involve three encoders that map input representations to a shared latent space $\mathcal{Z}$. Specifically, these are the image encoder $\mathbf{E}_I : \mathcal{I} \rightarrow \mathcal{Z}$, the textual encoder $\mathbf{E}_T : \mathcal{T} \rightarrow \mathcal{Z}$, the audio encoder $\mathbf{E}_A : \mathcal{A} \rightarrow \mathcal{Z}$ and the video encoder $\mathbf{E}_V : \mathcal{V} \rightarrow \mathcal{Z}$ for videos. Note that all seven elements employ only off-the-shelf frozen models.

Following the encouraging results of the preliminary analysis, MCANet utilizes θ_{LLM} and θ_C to compute the anomaly score for each frame. MCANet is designed to address limitations related to noise and lack of scene dynamics in frame-level captions by introducing three components: i) Image-Text Caption Cleaning through the vision-language representations of $\mathbf{E}_I$ and $\mathbf{E}_T$, ii) LLM-based Anomaly Scoring, encoding temporal information via θ_{LLM}, and iii) Video-Text Score Refinement of the anomaly scores via video-text similarity, using $\mathbf{E}_V$ and $\mathbf{E}_T$. Each component is described in detail in the following sections.

3.3 Image Captioning Branch

Image-Text Caption Cleaning. For each test video $\mathbf{V}$, θ_C is first employed to generate a caption $C_i^{(img)}$ for each frame $f_i \in \mathbf{V}$. Specifically, $\mathbf{C}^{(img)}$ denotes the sequence of captions, $\mathbf{C}^{(img)} = [C_1^{(img)}, \dots, C_M^{(img)}]$, where $C_i^{(img)} = \theta_C(f_i)$. However, as discussed in Sect. 3.2, the raw captions can be noisy, containing broken sentences or incorrect descriptions.

To address this issue, the entire set of captions $\mathbf{C}$ is used, under the assumption that within this set, there exist accurate and complete captions for the frames. This assumption is based on the typical scenario where a video features a scene captured by static cameras at a high frame rate, leading to overlapping semantic content among frames regardless of their temporal distances. Therefore, caption cleaning is treated as finding the semantically closest caption to a target frame f_i within $\mathbf{C}^{(img)}$. Formally, vision-language encoders are used to create a set of caption embeddings by encoding each caption in $\mathbf{C}^{(img)}$ via $\mathbf{E}_T$, i.e., $\{\mathbf{E}_T(C_1^{(img)}), \dots, \mathbf{E}_T(C_M^{(img)})\}$. For each frame $f_i \in \mathbf{V}$, its closest semantic caption is computed as:

$$\hat{C}_i^{(img)} = \arg \max_{C \in \mathbf{C}^{(img)}} \langle \mathbf{E}_I(f_i) \cdot \mathbf{E}_T(C^{(img)}) \rangle, \tag{2}$$

where $\langle \cdot, \cdot \rangle$ denotes cosine similarity, and $\mathbf{E}_I$ is the image encoder of the VLM. The cleaned set of captions is then constructed as $\hat{\mathbf{C}}^{(img)} = [\hat{C}_i^{(img)}, \dots, \hat{C}_M^{(img)}]$, replacing each initial caption $C_i^{(img)}$ with its counterpart $\hat{C}_i^{(img)}$ retrieved from $\mathbf{C}^{(img)}$. By performing this caption-cleaning process, captions of frames that are semantically more aligned with the visual content can be propagated, regardless of their temporal positioning, to improve or correct noisy descriptions.

3.4 Audio Captioning Branch

Audio-Text Caption Cleaning. For each test video $\mathbf{V}$, θ_A is first employed to generate a caption $C_i^{(aud)}$ for each audio segment $a_i \in \mathbf{V}$. Specifically,

$\mathbf{C}^{(aud)}$ denotes the sequence of captions, $\mathbf{C}^{(aud)} = [C_1^{(aud)}, \ldots, C_M^{(aud)}]$, where $C_i^{(aud)} = \theta_A(a_i)$. However, as discussed in Sect. 3.2, the raw captions can be noisy, containing broken sentences or incorrect descriptions.

To address this issue, the entire set of captions $\mathbf{C}$ is used, under the assumption that within this set, there exist accurate and complete captions for the audio segments. This assumption is based on the typical scenario where a video features a scene captured with continuous audio recording, leading to overlapping semantic content among audio segments regardless of their temporal distances. Therefore, caption cleaning is treated as finding the semantically closest caption to a target audio segment a_i within $\mathbf{C}^{(aud)}$. Formally, audio-language encoder is used to create a set of caption embeddings by encoding each caption in $\mathbf{C}^{(aud)}$ via $\mathbf{E}_T$, i.e., $\{\mathbf{E}_T(C_1^{(aud)}), \ldots, \mathbf{E}_T(C_M^{(aud)})\}$. For each audio segment $a_i \in \mathbf{V}$, its closest semantic caption is computed as:

$$\hat{C}_i^{(aud)} = \arg \max_{C \in \mathbf{C}^{(aud)}} \langle \mathbf{E}_A(a_i) \cdot \mathbf{E}_T(C^{(aud)}) \rangle, \tag{3}$$

where $\langle \cdot, \cdot \rangle$ denotes cosine similarity, and $\mathbf{E}_A$ is the audio encoder of the VLM. The cleaned set of captions is then constructed as $\hat{\mathbf{C}}^{(aud)} = [\hat{C}_i^{(aud)}, \ldots, \hat{C}_M^{(aud)}]$, replacing each initial caption $C_i^{(aud)}$ with its counterpart $\hat{C}_i^{(aud)}$ retrieved from $\mathbf{C}^{(aud)}$. By performing this caption-cleaning process, captions of audio segments that are semantically more aligned with the auditory content can be propagated, regardless of their temporal positioning, to improve or correct noisy descriptions.

3.5 LLM-Based Anomaly Scoring

The obtained caption sequence $\hat{\mathbf{C}}^{(img)}$, while cleaner than the initial set, lacks temporal information. To address this, the LLM is leveraged to provide temporal summaries. Specifically, a temporal window of T seconds, centered around f_i, is defined. Within this window, N frames are uniformly sampled, forming a video snippet $\mathbf{S}_i$, and a caption sub-sequence $\hat{\mathbf{C}}_i^{(img)} = \{\hat{C}_n^{(img)}\}_{n=1}^{N}$. The LLM is then queried with $\hat{\mathbf{C}}_i^{(img)}$ and a prompt P_S to get the temporal summary $\mathbf{TS}_i$ centered on frame f_i:

$$\mathbf{TS}_i = \theta_{LLM}(P_S \circ \hat{\mathbf{C}}_i^{(img)}) \circ \hat{\mathbf{C}}_i^{(aud)} \tag{4}$$

where the prompt P_S is formed as "Please summarize what happened in a few sentences, based on the following temporal description of a scene. Do not include any unnecessary details or descriptions."

Coupling Eq. (4) with the refinement process of Eq. (3), a textual description of the frame ($\mathbf{TS}_i$) which is semantically and temporally richer than $C_i^{(img)}$ and $C_i^{(aud)}$is obtained. With $\mathbf{TS}_i$, the LLM is queried for estimating an anomaly score. Following the same prompting strategy described in Sect. 3.2, the LLM is asked to assign to each temporal summary $\mathbf{TS}_i$ a score z_i in the interval $[0, 1]$. The score is obtained as:

$$z_i = \Phi_{LLM}(P_C \circ P_F \circ \mathbf{TS}_i) \tag{5}$$

where, as in Sect. 3.2, P_C is a context prompt containing VAD contextual priors, and P_F provides information on the desired output format.

3.6 Video-Text Score Refinement

By querying the LLM for each frame in the video with Eq. 5, the initial anomaly scores of the video $z = [z_1, \dots, z_M]$ are obtained. However, z is purely based on the language information encoded in their summaries, without considering the whole set of scores. To further refine them, visual information is leveraged to aggregate scores from semantically similar frames. Specifically, the video snippet $\mathbf{S}_i$ centered around f_i and all the temporal summaries are encoded using $\mathbf{E}_V$ and $\mathbf{E}_T$, respectively. Let $\mathcal{K}_i$ be the set of indices of the K-closest temporal summaries to $\mathbf{S}_i$ in $\{\mathbf{TS}_1, \dots, \mathbf{TS}_M\}$, where the similarity between $\mathbf{S}_i$ and a caption $\mathbf{TS}_j$ is the cosine similarity, i.e., $\langle \mathbf{E}_V(\mathbf{S}_i), \mathbf{E}_T(\mathbf{TS}_j)\rangle$. The refined anomaly score $\tilde{z}_i$ is obtained as:

$$\tilde{z}_i = \sum_{k \in \mathcal{K}_i} z_k \cdot \frac{e^{\langle \mathbf{E}_V(\mathbf{S}_i), \mathbf{E}_T(\mathbf{TS}_k)\rangle}}{\sum_{k \in \mathcal{K}_i} e^{\langle \mathbf{E}_V(\mathbf{S}_i), \mathbf{E}_T(\mathbf{TS}_k)\rangle}} \tag{6}$$

where $\langle \cdot, \cdot \rangle$ denotes the cosine similarity. Note that Eq. (6) applies the same principles as Eq. (3), refining frame-level estimations (i.e., score/captions) using their visual-language similarity (i.e., video/image) with other frames in the video. Finally, with the refined anomaly scores for the test video $\tilde{z} = [\tilde{z}_1, \dots, \tilde{z}_M]$, the anomalous temporal windows are identified via thresholding.

4 Experiments

This section provides a thorough evaluation of the MCANet framework. It begins with a description of the datasets and evaluation metrics used in the experiments, detailed in Sect. 4.1. The subsequent Sect. 4.2, outlines the implementation details of the approach. Qualitative results demonstrating the effectiveness of MCANet are presented in Sect. 4.3. A detailed comparison with state-of-the-art methods is presented in Sect. 4.4, followed by ablation studies in Sect. 4.5, which assess the impact of each component of MCANet.

4.1 Datasets and Evaluation Metrics

Datasets: The experiments are conducted on two benchmark large-scale datasets, namely UCF-Crime [7] and XD-Violence [22]. The UCF-Crime dataset is a large-scale collection comprising 1,900 untrimmed real-world surveillance videos. It includes both outdoor and indoor environments, offering a total duration of 128 h. The dataset is divided into 13 distinct classes of anomalous events such as fighting, stealing, assault etc. XD-Violence is a large-scale multimodal dataset that includes audio signals encompassing 4754 movies and YouTube videos. The total duration of this dataset is 217 h. It includes 6 distinct classes of anomalous events such as car accident, riot, explosion etc.

Evaluation Metrics: To evaluate the anomaly detection performance, the Area Under the Curve (AUC) of the frame-level receiver operating characteristics (ROC) is used as the main evaluation metric for the UCF-Crime dataset. Following [23], this evaluation calculates AUC for the entire test set, denoted as AUC. Additionally, the AUC is computed specifically for abnormal videos, referred to as Ano-AUC. This approach excludes normal videos where all clips are labeled as normal (label 0), retaining only the abnormal ones containing both normal and abnormal clips (labels 0 and 1). Moreover, the AUC of the frame-level precision-recall curve (AP) is utilized for the XD-Violence dataset.

4.2 Implementation Details

The experiments are conducted on NVIDIA RTX 4070 GPU using the PyTorch framework. For the optimization of the network, Adam is leveraged as the optimizer with a batch size of 128. On the XD-Violence dataset, the learning rate and total epoch are set as 5×10^{-4} and 50, respectively, and on the UCF-Crime dataset, the learning rate and total epoch are set as 3×10^{-4} and 50, respectively. In this study, the ViT in the LanguageBind model [24] serves as the Video Encoder, the pre-trained ViT L/14 [25] is used as the image encoder, and the VGGish network [26] is employed as the audio encoder. FUSECAP [21] is used as vision-language model, while Pengi [27] is used as audio-language model that serves as caption generators from visual frames and audio signals respectively. Additionally, the pre-trained LLaMA from Video-LLaMA [18] is implemented to retain general video description. For computational efficiency, each video is randomly sampled at 16-frame intervals.

4.3 Qualitative Results

The frame-wise AUC results of sample test videos of UCF-Crime and XD-Violence dataset are visualized in Fig. 3. Figure 3(a) and (b) depicts the qualitative results of Vandalism and Robbery anomaly from the UCF-Crime dataset. Figure 3(a) and (b) shows the qualitative results of Vandalism and Robbery anomaly from the XD-Violence dataset. It has been observed that the temporal summaries generated by LLMs accurately capture the visual content of anomalous situations. This depiction empowers MCANet to effectively identify the anomalies, ensuring accurate detection of the abnormal events.

4.4 Quantitative Results

The proposed MCANet is compared with state-of-the-art(SOTA) approaches, including one-class approaches [37–39], unsupervised approaches [4,6,42,43], weakly-supervised approaches [7,23,28,29,31], training-free approaches [24,44–46]. The performance comparison on the UCF-Crime dataset is presented in Table 1. Remarkably, the proposed MCANet achieves better performance compared to both one-class and unsupervised baselines. It achieves a higher AUC,

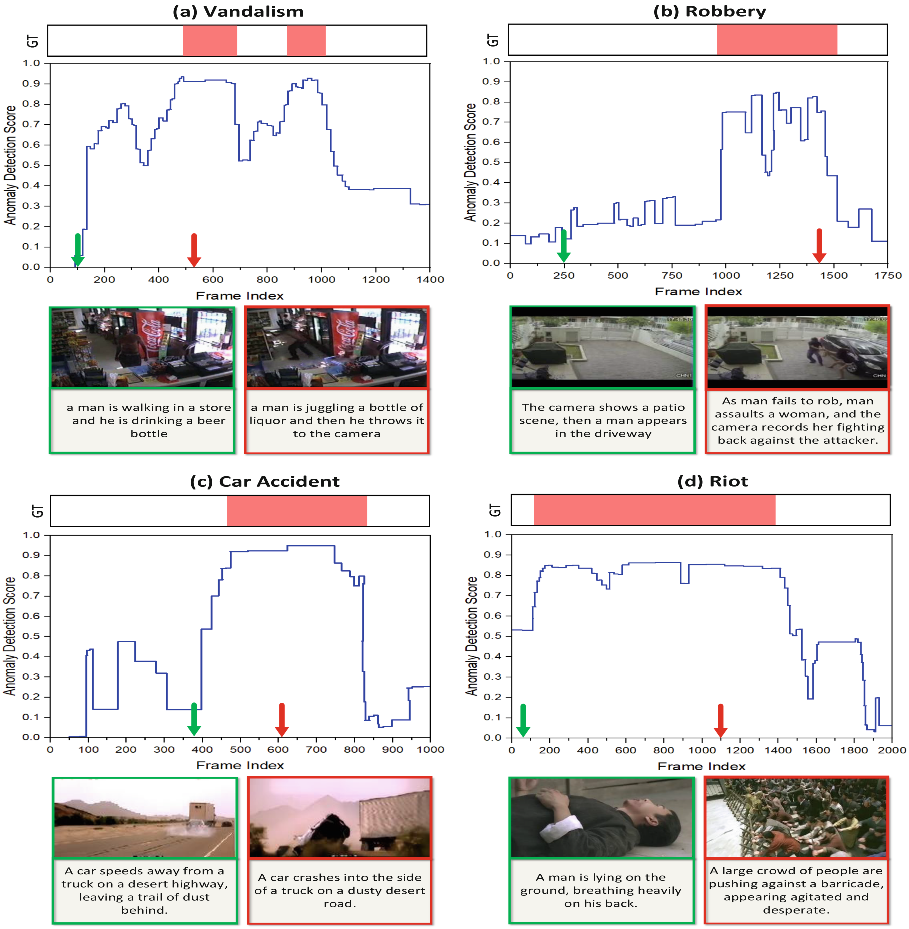


Fig. 3. Qualitative results of the proposed method on UCF-Crime and XD-Violence dataset. Figure 3(a) and Fig. 3(b) are from UCF-Crime. Figure 3(c) and Fig. 3(d) are from the XD-Violence dataset. Blue curves show anomaly scores of our method. Light Pink areas indicate the ground-truth abnormal frames. Each red and green box shows the abnormal and normal event. Best viewed in color. (Color figure online)

with a significant improvement of +9.77% when compared to SACR [40] and a minor improvement of +1.62% against the current state of the art obtained by C2FPL [6]. Compared to training-free approaches, the proposed MCANet achieves highest AUC of 82.47% and The performance comparison on the XD-Violence dataset is reported in Table 2. Consistent with the results on UCF-Crime, the proposed method outperforms all one-class [37–39], unsupervised approaches [4,6,42,43] and training-free approaches [24,44–46] by a significant margin in Average Precision (AP). For example, using I3D features, MCANet performs better than BODS [39] by 13.08%, GODS [39] by +10.88%, DyAn-

Table 1. Performance comparison on UCF-Crime

Supervision	Method	Features	Explanation	AUC(%)	Ano-AUC(%)
Weak	Deep-MIL [7]	C3D	✗	75.42	54.25
	HL-Net [28]	I3D	✗	82.45	60.27
	MS-BSAD [29]	I3D	✗	83.54	–
	MSDeepNet [30]	I3D	✗	84.72	–
	NG-MIL [31]	I3D	✗	85.63	–
	RTFM [32]	I3D	✗	85.66	63.86
	AnomalyCLIP [33]	ViT	✗	86.36	–
	MGFN[34]	VideoSwin	✗	86.67	–
	DMU [35]	I3D	✗	86.75	66.8
	UB-MIL [23]	X-CLIP	✗	86.97	68.94
	CLIP-TSA [36]	CLIP	✗	87.58	69.31
One-Class	SVM baseline [7]	–	✗	50.10	–
	Hasan et al. [37]	–	✗	51.20	39.43
	SSV [38]	–	✗	58.50	–
	BODS[39]	I3D	✗	68.26	–
	GODS[39]	I3D	✗	70.46	–
	SACR[40]	–	✗	72.70	–
Un	Tur et al.[41]	ResNet	✗	65.22	–
	Tur et al.[42]	ResNet	✗	66.85	–
	Zaheer et al.[4]	ResNext	✗	74.20	–
	DyAnNet[43]	I3D	✗	79.76	–
	C2FPL[6]	I3D	✗	80.85	–
Training-Free	ZS-CLIP[44]	ViT	✓	53.16	48.67
	ZS IMAGEBIND (IMAGE) [24]	ViT	✓	53.65	50.65
	ZS IMAGEBIND (VIDEO) [24]	ViT	✓	55.78	52.93
	LLAVA-1.5 [45]	ViT	✓	72.84	62.14
	LAVAD [46]	ViT	✓	80.28	63.21
Training-Free	**MCANet (Ours)**	**I3D**	✓	**81.34**	**65.78**
	MCANet (Ours)	**ViT**	✓	**82.47**	**67.12**

Net [43] by +1.58%, C2FPL [6] by +0.49%. However, the proposed training-free approach fails to achieve satisfactory performance compared to recent weakly-supervised approaches. This performance gap is due to the lack of specific visual priors and the limitations of vision-language models (VLMs) and audio language models (ALMs) in focusing predominantly on prominent foreground signals rather than contextually relevant background information.

Table 2. Performance Comparison on XD-Violence

Supervision	Method	Features	Explanation	AUC(%)	AP(%)
Weak	Deep-MIL [7]	C3D	✗	–	75.18
	HL-Net [28]	I3D	✗	–	78.10
	MS-BSAD [29]	I3D	✗	–	78.92
	NG-MIL [31]	I3D	✗	–	78.51
	RTFM [32]	I3D	✗	–	78.27
	AnomalyCLIP [33]	ViT	✗	–	78.55
	MGFN [34]	I3D	✗	–	79.19
	MGFN [34]	VideoSwin	✗	–	80.11
	DMU [35]	I3D	✗	–	82.41
	CLIP-TSA [36]	CLIP	✗	–	82.17
One-Class	Hasan et al. [37]	AE^{RGB}	✗	50.32	–
	Lu et al. [47]	Dictionary	✗	53.56	–
	BODS [39]	I3D	✗	57.32	–
	GODS [39]	I3D	✗	61.56	–
Un	RareAnom [48]	I3D	✗	68.33	–
	C2FPL [6]	I3D	✗	80.09	–
Training-Free	ZS-CLIP [44]	ViT	✓	38.21	17.83
	ZS IMAGEBIND (IMAGE) [24]	ViT	✓	58.81	27.25
	ZS IMAGEBIND (VIDEO) [24]	ViT	✓	55.06	25.36
	LLAVA-1.5 [45]	ViT	✓	79.62	50.26
	LAVAD [46]	ViT	✓	85.36	62.01
Training-Free	**MCANet (Ours)**	**I3D**	✓	**86.81**	**68.19**
	MCANet (Ours)	**ViT**	✓	**87.43**	**69.72**

4.5 Ablation Studies

The ablations were carried out only on the UCF-Crime dataset. Initially, the effectiveness of each proposed component of MCANet was evaluated. Subsequently, the influence of task priors in the context prompt on the estimation of anomaly scores was analyzed.

Impact of Each Component of the Proposed MCANet. Experiments are conducted to investigate the impact of each component of the proposed MCANet. As shown in Table 3, The first row shows the results when both Image-Text Caption Cleaning and Audio-Text Caption Cleaning components are omitted, resulting in a significant degradation in performance, with an AUC of 78.53%. In the second row, when only the Audio-Text Caption Cleaning component is omitted, the AUC drops to 75.31%, demonstrating a substantial decrease of

7.16% compared to the full model (row 5). The third row excludes LLM-based Anomaly Scoring, relying solely on cleaned captions and video-text score refinement, leading to an AUC of 74.70%. The significant drop of 7.77% in AUC highlights the importance of the LLM's role in summarizing temporal information and accurately estimating anomaly scores. In the fourth row, when Video-Text Score Refinement is not used, the AUC decreases to 75.79%, showing a 6.68% reduction compared to the full MCANet model. This result confirms the critical contribution of score refinement, which aggregates scores from semantically similar frames to enhance the overall VAD accuracy. The final row shows the complete MCANet model with all components included, achieving the highest AUC of 82.47%. This demonstrates the cumulative effectiveness of integrating Image-Text Caption Cleaning, Audio-Text Caption Cleaning, LLM-based Anomaly Scoring, and Video-Text Score Refinement in boosting VAD performance.

Table 3. Impact of each proposed component on the UCF-Crime Dataset.

Image-Text Caption Cleaning	Audio-Text Caption Cleaning	LLM-based Anomaly Scoring	Video-Text Score Refinement	AUC(%)
✗	✗	✓	✓	78.53
✓	✗	✗	✓	75.31
✗	✓	✗	✓	74.70
✓	✓	✓	✗	75.79
✓	✓	✓	✓	**82.47**

Impact of Task Priors in the Context Prompt. The impact of task priors in the context prompt was examined, and the results are presented in Table 4. Specifically, two types of priors were investigated: anomaly prior and impersonation. The anomaly prior involves guiding the LLM with context related to anomalies, such as criminal activities, which could enhance the relevance of the semantic context. Impersonation, on the other hand, allows the LLM to process the input from the viewpoint of potential end-users of a VAD system, such as law enforcement agencies. The ablation studies begin with a base context prompt that does not include any specific priors: "How would you rate the event described on a scale from 0 to 1, where 0 represents normal and 1 represents anomalous behavior?" (Row 1). When only the anomaly prior is added to this prompt (Row 2), there is a slight improvement in the AUC, reaching 81.79%, suggesting that the anomaly prior does contribute positively, but not drastically, to the performance of LLM. When the impersonation prior is used on its own (Row 3), the AUC increases to 80.82%, indicating that having the LLM adopt the perspective of a law enforcement agency improves its ability to detect anomalies, though still not as effectively as the full context. Finally, incorporating both the anomaly prior and the impersonation (Row 4) results in the highest AUC of 82.47%, demonstrating the combined benefits of these priors in enhancing the performance of LLM for anomaly detection. This suggests that both priors are complementary, with their integration providing a more robust framework for the LLM to assess anomalies in video data.

Table 4. Impact of task priors in the context prompt when querying for anomaly scores.

Anomaly Prior	Impersonation	AUC(%)
✗	✗	80.48
✗	✓	80.82
✓	✗	81.79
✓	✓	**82.47**

5 Conclusion

To address training-free video anomaly detection (VAD), a novel framework called MCANet is introduced. MCANet identifies anomalies in video sequences without prior domain knowledge by leveraging off-the-shelf vision-language model (VLM), audio-language model (ALM) and large language model (LLM). It dynamically generates and analyzes textual and audio-visual descriptions of video frames, using an innovative image-text and audio-text caption cleaning module. These descriptions are processed through a prompting mechanism for LLMs to perform temporal aggregation and anomaly score estimation. Experimental results on two large-scale benchmark datasets demonstrate that MCANet outperforms existing unsupervised and one-class approaches without any training or data collection. However, the performance of the MCANet heavily depends on the quality of pre-trained models and the effectiveness of prompting strategies, highlighting areas for further research and community involvement.

References

1. Zhao, M., Liu, Y., Liu, J., Zeng, X.: Exploiting spatial-temporal correlations for video anomaly detection. In: 2022 26th International Conference on Pattern Recognition (ICPR), pp. 1727–1733. IEEE (2022)
2. Lee, J., Nam, W.-J., Lee, S.-W.: Multi-contextual predictions with vision transformer for video anomaly detection. In: 2022 26th International Conference on Pattern Recognition (ICPR), pp. 1012–1018. IEEE (2022)
3. Deng, H., Zhang, Z., Zou, S., Li, X.: Bi-directional frame interpolation for unsupervised video anomaly detection. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pp. 2634–2643 (2023)

4. Zaheer, M.Z., Mahmood, A., Khan, M.H., Segu, M., Yu, F., Lee, S.-I.: Generative cooperative learning for unsupervised video anomaly detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 14744–14754 (2022)
5. Sun, Z., Wang, P., Zheng, W., Zhang, M.: Dual GroupGAN: an unsupervised four-competitor (2V2) approach for video anomaly detection. Pattern Recogn. **153**, 110500 (2024)
6. Al-lahham, A., Tastan, N., Zaheer, M.Z., Nandakumar, K.: A coarse-to-fine pseudo-labeling (C2FPL) framework for unsupervised video anomaly detection. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pp. 6793–6802 (2024)
7. Sultani, W., Chen, C., Shah, M.: Real-world anomaly detection in surveillance videos. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 6479–6488 (2018)
8. Huang, C., et al.: Weakly supervised video anomaly detection via self-guided temporal discriminative transformer. IEEE Trans. Cybern. **54**(5), 3197–3210 (2022)
9. Ullah, W., Ullah, F.U.M., Khan, Z.A., Baik, S.W.: Sequential attention mechanism for weakly supervised video anomaly detection. Expert Syst. Appl. **230**, 120599 (2023)
10. Karim, H., Doshi, K., Yilmaz, Y.: Real-time weakly supervised video anomaly detection. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 6848–6856 (2024)
11. Yan, L., Han, C., Xu, Z., Liu, D., Wang, Q.: Prompt learns prompt: exploring knowledge-aware generative prompt collaboration for video captioning. In: Proceedings of International Joint Conference on Artificial Intelligence (IJCAI), pp. 1622–1630 (2023)
12. Lin, K., et al.: MM-VID: Advancing video understanding with GPT-4v (ision). arXiv preprint arXiv:2310.19773 (2023)
13. Chen, G., et al.: VideoLLM: Modeling video sequence with large language models. arXiv preprint arXiv:2305.13292 (2023)
14. Jiang, C., et al.: BUS: efficient and effective vision-language pre-training with bottom-up patch summarization. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 2900–2910 (2023)
15. Li, X., et al.: Oscar: object-semantics aligned pre-training for vision-language tasks. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J.-M. (eds.) ECCV 2020. LNCS, vol. 12375, pp. 121–137. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-58577-8_8
16. Bain, M., Nagrani, A., Varol, G., Zisserman, A.: Frozen in time: a joint video and image encoder for end-to-end retrieval. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 1728–1738 (2021)
17. Li, K., et al.: VideoChat: Chat-centric video understanding. arXiv preprint arXiv:2305.06355 (2023)
18. Zhang, H., Li, X., Bing, L.: Video-LLaMa: An instruction-tuned audio-visual language model for video understanding. arXiv preprint arXiv:2306.02858 (2023)
19. Lin, B., Zhu, B., Ye, Y., Ning, M., Jin, P., Yuan, L.: Video-LLaVA: Learning united visual representation by alignment before projection. arXiv preprint arXiv:2311.10122 (2023)
20. He, B., et al.: MA-LMM: Memory-augmented large multimodal model for long-term video understanding. arXiv preprint arXiv:2404.05726 (2024)

21. Rotstein, N., Bensaïd, D., Brody, S., Ganz, R., Kimmel, R.: FuseCap: leveraging large language models for enriched fused image captions. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pp. 5689–5700 (2024)
22. Wu, P., et al.: Not only look, but also listen: learning multimodal violence detection under weak supervision. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J.-M. (eds.) ECCV 2020. LNCS, vol. 12375, pp. 322–339. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-58577-8_20
23. Lv, H., Yue, Z., Sun, Q., Luo, B., Cui, Z., Zhang, H.: Unbiased multiple instance learning for weakly supervised video anomaly detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 8022–8031 (2023)
24. Girdhar, R., et al.: ImageBind: one embedding space to bind them all. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 15180–15190 (2023)
25. Dosovitskiy, A., et al.: An image is worth 16×16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
26. Hershey, S., et al.: CNN architectures for large-scale audio classification. In: 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 131–135. IEEE (2017)
27. Deshmukh, S., Elizalde, B., Singh, R., Wang, H.: Pengi: an audio language model for audio tasks. Adv. Neural. Inf. Process. Syst. **36**, 18090–18108 (2023)
28. Wu, P., Liu, X., Liu, J.: Weakly supervised audio-visual violence detection. IEEE Transactions on Multimedia (2022)
29. Zhen, Y., Guo, Y., Wei, J., Bao, X., Huang, D.: Multi-scale background suppression anomaly detection in surveillance videos. In: 2021 IEEE International Conference on Image Processing (ICIP), pp. 1114–1118. IEEE (2021)
30. Dev, P.P., Das, P., Hazari, R.: MSDeepNet: a novel multi-stream deep neural network for real-world anomaly detection in surveillance videos. In: International Conference on Deep Learning Theory and Applications, pp. 157–172. Springer (2023)
31. Park, S., Kim, H., Kim, M., Kim, D., Sohn, K.: Normality guided multiple instance learning for weakly supervised video anomaly detection. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pp. 2665–2674 (2023)
32. Tian, Y., Pang, G., Chen, Y., Singh, R., Verjans, J.W., Carneiro, G.: Weakly-supervised video anomaly detection with robust temporal feature magnitude learning. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 4975–4986 (2021)
33. Zanella, L., Liberatori, B., Menapace, W., Poiesi, F., Wang, Y., Ricci, E.: Delving into clip latent space for video anomaly recognition. arXiv preprint arXiv:2310.02835 (2023)
34. Chen, Y., Liu, Z., Zhang, B., Fok, W., Qi, X., Yik-Chung, W.: MGFN: magnitude-contrastive glance-and-focus network for weakly-supervised video anomaly detection. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 37, pp. 387–395 (2023)
35. Zhou, H., Junqing, Yu., Yang, W.: Dual memory units with uncertainty regulation for weakly supervised video anomaly detection. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 37, pp. 3769–3777 (2023)
36. Joo, H.K., Vo, K., Yamazaki, K., Le, N.: CLIP-TSA: clip-assisted temporal self-attention for weakly-supervised video anomaly detection. In: 2023 IEEE International Conference on Image Processing (ICIP), pp. 3230–3234. IEEE (2023)

37. Hasan, M., Choi, J., Neumann, J., Roy-Chowdhury, A.K., Davis, L.S.: Learning temporal regularity in video sequences. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 733–742 (2016)
38. Sohrab, F., Raitoharju, J., Gabbouj, M., Iosifidis, A.: Subspace support vector data description. In: 2018 24th International Conference on Pattern Recognition (ICPR), pp. 722–727. IEEE (2018)
39. Wang, J., Cherian, A.: GODS: generalized one-class discriminative subspaces for anomaly detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 8201–8211 (2019)
40. Sun, C., Jia, Y., Hu, Y., Wu, Y.: Scene-aware context reasoning for unsupervised abnormal event detection in videos. In: Proceedings of the 28th ACM International Conference on Multimedia, pp. 184–192 (2020)
41. Tur, A.O., Dall'Asen, N., Beyan, C., Ricci, E.: Exploring diffusion models for unsupervised video anomaly detection. In: 2023 IEEE International Conference on Image Processing (ICIP), pp. 2540–2544. IEEE (2023)
42. Tur, A.O., Dall'Asen, N., Beyan, C., Ricci, E.: Unsupervised video anomaly detection with diffusion models conditioned on compact motion representations. In: International Conference on Image Analysis and Processing, pp. 49–62. Springer (2023)
43. Thakare, K.V., Raghuwanshi, Y., Dogra, D.P., Choi, H., Kim, I.-J.: DyAnNet: a scene dynamicity guided self-trained video anomaly detection network. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pp. 5541–5550 (2023)
44. Radford, A., et al.: Learning transferable visual models from natural language supervision. In: International Conference on Machine Learning, pp. 8748–8763. PMLR (2021)
45. Zhaopeng, G., Zhu, B., Zhu, G., Chen, Y., Tang, M., Wang, J.: AnomalyGPT: detecting industrial anomalies using large vision-language models. In Proceedings of the AAAI Conference on Artificial Intelligence, vol. 38, pp. 1932–1940 (2024)
46. Zanella, L., Menapace, W., Mancini, M., Wang, Y., Ricci, E.: Harnessing large language models for training-free video anomaly detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 18527–18536 (2024)
47. Lu, C., Shi, J., Jia, J.: Abnormal event detection at 150 FPS in MATLAB. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 2720–2727 (2013)
48. Thakare, K.V., Dogra, D.P., Choi, H., Kim, H., Kim, I.-J.: RareAnom: a benchmark video dataset for rare type anomalies. Pattern Recog. **140**, 109567 (2023)

2by2: Weakly-Supervised Learning for Global Action Segmentation

Elena Bueno-Benito(✉) and Mariella Dimiccoli

Institut de Robòtica i Informàtica Industrial, CSIC-UPC, Llorens i Artigas 4-6, 08028
Barcelona, Spain
{ebueno,mdimiccoli}@iri.upc.edu

Abstract. This paper presents a simple yet effective approach for the poorly investigated task of global action segmentation, aiming at grouping frames capturing the same action across videos of different activities. Unlike the case of videos depicting all the same activity, the temporal order of actions is not roughly shared among all videos, making the task even more challenging. We propose to use activity labels to learn, in a weakly-supervised fashion, action representations suitable for global action segmentation. For this purpose, we introduce a triadic learning approach for video pairs, to ensure intra-video action discrimination, as well as inter-video and inter-activity action association. For the backbone architecture, we use a Siamese network based on sparse transformers that takes as input video pairs and determine whether they belong to the same activity. The proposed approach is validated on two challenging benchmark datasets: Breakfast and YouTube Instructions, outperforming state-of-the-art methods.

Keywords: Temporal Action Segmentation · Weakly-Supervised Learning · Video Alignment

1 Introduction

Action segmentation, the task of classifying each frame of an untrimmed video plays a fundamental role in various applications such as video surveillance, sports analysis, and content-based video retrieval [21,50]. Recently, this task has received significant attention from the research community. The most reliable approaches for action segmentation are fully supervised methods, which require expensive data annotations [5,6,19,27,32,48]. The need for more scalable and practical solutions has led to an increasing interest in developing weakly-supervised [9,30,31,33,41,46,49] and unsupervised techniques [7,11,12,14,23,24,26,28,35,37,42,43,45,47].

Weakly-supervised methods learn to partition videos into action segments using only transcript annotations for each video, typically in the form of actions transcripts (ordered lists of actions) or action sets (unique actions derived from narrations, captions or meta-tags) [31,41,46,49]. This weakly-supervised paradigm contrasts with unsupervised methods, broadly categorized into three types, depending

A. Antonacopoulos et al. (Eds.): ICPR 2024, LNCS 15332, pp. 376–391, 2025.
https://doi.org/10.1007/978-3-031-78125-4_26

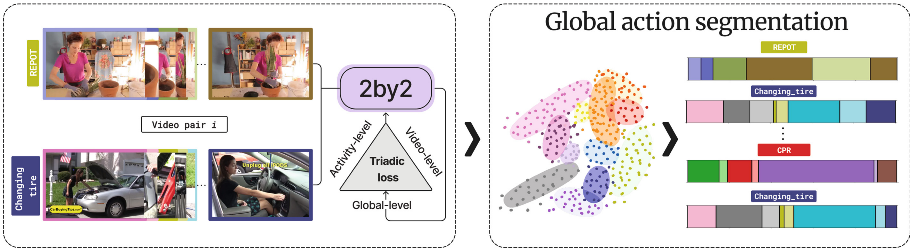


Fig. 1. Our approach compares video pairs through a Siamese network by using binary labels indicating if the videos belong to the same activity or not. We propose a triadic loss function modelling intra-video discrimination, inter-video and inter-activity associations for clustering actions across videos of different activities.

on the matching objective [13]: video-level, activity-level, and global-level. Video-level segmentation methods aim to segment a single video sequence into distinct actions without considering the relationships between actions in different videos [7,16,28,35,47]. While they can be effective for practical applications requiring to segment isolated videos one by one, they fail to generalize actions across different videos. Instead, activity-level segmentation methods focus on matching actions across videos that depict the same complex activity [14,23,24,26,42,46]. These methods generally perform poorly at video-level unless temporal smoothing within segments is explicitly modelled. In addition, as they assume or estimate a transcript for each video or set of videos belonging to the same activity, their generalization ability to other activities is hampered. Only Ding *et al.*[14] directly addressed global-level segmentation railing on complex activity labels to help discover the constituent actions; however, they do not explicitly model the alignment of actions across videos of the same activity.

In this paper, we propose a strategy to discover actions across various complex activity videos, offering a broader and more generalized understanding of actions. Our approach does not require knowledge of video transcripts, but only binary labels indicating whether each pair of videos belongs to the same activity. Therefore, as a weakly-supervised method, it occupies a unique position in the spectrum of action segmentation methods.

Our solution, depicted in Fig. 1, aims to enhance the clustering of actions in videos on a global scale through the implementation of a Siamese network based on transformers. This network is designed to address the task of determining whether two videos depict the same activity. Instead of using a standard cross-entropy loss, we propose a triadic loss function capturing the temporal dynamics within individual videos, between similar videos, and across various activities. Our contributions are as follows:

1. We propose a novel weakly-supervised framework for the task of global action segmentation that relies on binary activity labels to discover action clusters across videos of different activities.

2. We introduce a transformer-based Siamese architecture that takes as input pairs of videos, determines if they belong to the same activity or not and aligns them temporally if predicts that they depict the same activity.
3. We introduce a triadic loss function that models intra-video action discrimination at the video-level, inter-video and inter-activity action associations at activity and global-level respectively, for robust action understanding.
4. We achieve state-of-the-art results on the *Breakfast (BF)* and *Inria Instructional Videos (YTI)* benchmark datasets, demonstrating the method's effectiveness and generalization ability across activities.

2 Related Work

2.1 Action Segmentation

For a comprehensive and recent survey on temporal action segmentation tasks, readers are referred to [13].

Supervised Action Segmentation. Supervised approaches have seen significant advancements over recent years [5,6,19,27,32,48]. Recently, UVAST [6] integrates fully and timestamp-supervised learning paradigms via sequence-to-sequence translation. This method refines predictions by aligning frame labels with predicted action sequences. LTContext [5] iterates between windowed local attention and sparse long-term context attention, effectively balancing computational complexity and segmentation accuracy. Lastly, FACT [32] performs temporal modelling at both frame-level and action-level, facilitating bidirectional information transfer and iterative feature refinement. However, being fully supervised, all these methods are not scalable and not suited for real applications.

Weakly-Supervised Action Segmentation. Weakly-supervised techniques have been developed to reduce the need for large annotated datasets. These approaches typically learn to partition a video into several action segments from training videos only using transcripts or other human-generated information to generate pseudo-labels for training [30,31,33,41,46,49]. Transcripts have been shown to outperform action set-based methods, while timestamp-based approaches achieve the best results. This suggests that higher levels of supervision generally lead to better performance. In recent years, DP-DTW [9] has advanced weakly-supervised segmentation by training class-specific discriminative action prototypes. This method represents videos by concatenating prototypes based on transcripts and improves inter-class distinction through discriminative losses. Some methods leverage machine learning models to infer video segments, such as TASL [30]. Recently, more efficient alignment-free methods have been proposed. MuCon [41] learns from the mutual consistency between two forms of segmentation: framewise classification and category/length pairs. POC [31] introduces a loss function to ensure the output order of any two actions aligns with the transcript. Conversely, ATBA [46] propose an approach that incorporates alignment by directly localizing action transitions for efficient pseudo-segmentation generation during training, eliminating the need for time-consuming frame-by-frame alignment. None of these methods explicitly addresses the problem of global action segmentation.

Unsupervised Action Segmentation. Unsupervised approaches have been explored by several studies to eliminate the need for annotations [1,7,11,14,16,23,24,26,28,35–37,42,43,45,47]. As the estimated clusters, lack of semantic labels, the evaluation process requires finding the Hungarian correspondence between the clusters and the actual action classes. The Hungarian matching can be performed for video-level segmentation [1,7,16,28,35,36], activity-level segmentation [14,23,24,26,37,42,43,45,47], or for a global scope across an entire set of videos [14,23,26]. Depending on the hierarchical level used, methods aim to improve segmentation through these correspondences. Unsupervised techniques in action segmentation typically involve a two-step process: first, learning action representations in a self-supervised manner, followed by employing clustering algorithms to perform action segmentation, assuming a prior knowledge of the number of clusters.

In the realm of video-level action segmentation, LSTM+AL [1] introduced a novel self-supervised methodology for real-time action boundary detection. Furthermore, it is worth noticing that clustering approaches based on specific similarity metrics have been relatively under-explored in the field of action segmentation. One such method is TW-FINCH [35], which captures spatio-temporal similarities among video frames. This employs a temporally weighted hierarchical clustering algorithm, grouping video frames without the need for extensive pre-training, as it directly operates on pre-computed features that augment the conventional FINCH approach with temporal considerations [36]. In a similar vein, ABD [16] identifies action boundaries by detecting abrupt change points along the similarity chain between consecutive features.

Action representation learning at the individual video level has also gained interest. TSA [7] proposed a method that focuses on this aspect, employing a shallow neural network trained with a triplet loss and a novel triplet selection strategy to learn action representations. These learned representations can be processed using generic clustering algorithms to obtain segmentation outputs. Lastly, the OTAS framework has emerged, offering an unsupervised boundary detection method that combines global visual features, local interacting features, and human-object relational features, contributing to the evolving landscape of action segmentation techniques [28].

Some approaches at the activity-level leverage the order of scripted activities, emphasizing the minimization of prediction errors, like CTE [23]. Other works combined temporal embedding with visual encoder-decoder pipelines with visual reconstruction loss [43] or with discriminative embedding loss [40]. ASAL [26] explored deep learning architectures, such as ensembles of autoencoders and classification networks that exploit the relationship between actions and activities. CAD [14] introduced a framework that discovers global action prototypes based on high-level activity labels. One notable aspect of these methods is the recognition that actions in task-oriented videos tend to occur in similar temporal contexts. As a result, strong temporal regularization techniques have been developed to partially obscure visual similarities [23,37]. Recently, optimal transport has gained popularity in unsupervised learning to generate effective pseudo-labels and train for frame-level action classification. TOT [24] proposed a joint self-supervised representation learning and online clustering approach that directly optimizes unsupervised activity segmentation using video frame clustering as a pretext task. UFSA [42] extends TOT by combining frame and segment-level cues to improve permutation-aware activity segmentation. Furthermore, TOT and UFSA use a Hidden Markov Model (HMM) approach to decode segmentations given a fixed or estimated action order, respectively. In contrast, ASOT [47] proposed a method via optimal transport that yields temporally consistent segmentations without prior knowledge of the action ordering, required by previous approaches. Suitable for both pseudo-labeling and decoding.

Although global-level understanding provides the most comprehensive insight into the relationships between activities and actions in videos, only a few methods have explored training at this level. CAD [14] is the first work to operate at the highest level of global matching. In CTE [23], the methods extended their configuration considering all complex activities. Firstly, the protocol executes a bag-of-words clustering on the videos to divide them into multiple pseudo-activities. Subsequently, they perform action clustering within each pseudo-activity individually. In other words, they apply their action segmentation at the activity level within classes of pseudo-activity. Their approach still does not accommodate potential actions shared between activities. ASAL [26] and CAD [14] present their results aligned with this protocol.

2.2 Video Alignment

Video alignment is a process aimed at synchronizing and matching video sequences for various applications, such as action recognition model creation, behavioural analysis, and multimedia content generation. This field encompasses a range of techniques. Traditionally, methods like Dynamic Time Warping (DTW), Canonical Correlation Analysis, ranking or match-classification objectives, and the differentiable version of DTW, Soft-DTW, have been used to tackle the challenging task of aligning video frames [3,4,10,38] in videos depicting a same action. Recently, LAV [20] have utilized Soft-DTW combined with temporal intra-video contrastive loss to align video frames effectively. Drop-DTW [17], an extension of DTW, introduces a "trash bucket" to the cost matrix, allowing for the classification of background frames and robust alignment in the presence of outliers. VAVA [29] employs optimal transport with a bi-modal Gaussian prior and a virtual frame for unmatched frames.

TCC [18] was the first to introduce cycle-consistency for aligning video frames by maximizing cycle-consistent embeddings between sequences. GTCC [15] extends the TCC approach to manage more complex alignment scenarios. However, most of these techniques were developed for general video alignment or related tasks, and their direct application to unsupervised action segmentation has been never explored so far. In this paper, we propose for the first time to leverage video alignment for action segmentation.

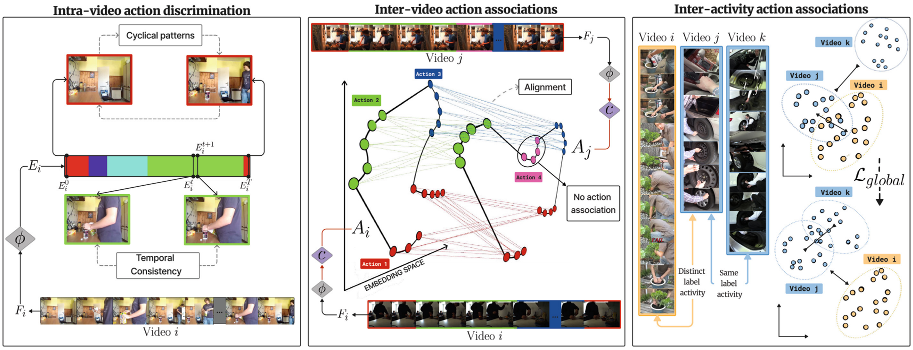


Fig. 2. Overview of the proposed 2by2 framework. The figure illustrates our triadic learning approach: intra-video action discrimination, which enhances cross-temporal consistency within a single video (first box); inter-video action associations, which align action frames among similar videos (second box); and inter-activity action associations, which establish global correspondence between different videos (third box). The red arrows indicate steps specific to the training phase. (Color figure online)

3 2by2: Learning Unknown Actions in a Global Manner

This section presents a weakly-supervised, triadic action learning approach for global action segmentation (see Fig. 2), aiming at modeling:

(i) *Intra-video action discrimination* (video level): Video frames sharing the same action with their nearest neighbours exhibit temporal consistency. Moreover, actions typically do not occur at the beginning or end of videos. Thus, a video can be interpreted as a cyclic temporal sequence.
(ii) *Inter-video action associations* (activity level): For videos categorized under the same activity, segments within these videos exhibit similarity, facilitating the alignment of actions across them.
(iii) *Inter-activity action associations* (global level): Videos representing different activities that share common actions should be closer in the representational space compared to those that do not share actions.

3.1 Problem Formulation

Given a large set V of complex activity videos from a dataset belonging to C complex activities, each video v_i in V is annotated with a complex activity label $a \in [1, C]$. Our objective is to associate each video frame x_t, with an action label n from N possible actions. These N actions are constituent steps shared among the C complex activities. For each video v_i, we define the feature matrix F_i, where each row F_i^t corresponds to an d-dimensional feature vector at time t in a video v_i. Given the initial features of a video F_i, our objective is to learn a parametric function ϕ that categorizes video frames into the N possible actions, resulting in embeddings E_i, obtained as $E_i = \phi(F_i), \forall v_i \in V$.

3.2 Architecture

To learn ϕ, we propose a Siamese architecture that takes as input pairs (F_i, F_j) for all $v_i, v_j \in V$ with $i \neq j$. This architecture consists of two identical LTContext networks [5], specifically designed to capture long-term temporal dependencies, that work in tandem and compare the similarity between their outputs, denoted as $(E_i,\ E_j)$ at the end.

During training, to ensure that videos sharing the same activity have well-aligned representations, we introduce a context-drop function c, inspired by [15]. This function is designed to handle background and redundant frames by enforcing multi-cycle consistency for alignable embeddings and poor alignment for droppable embeddings. The context-adjusted embeddings are calculated as $A_i = c(E_i), \forall v_i \in V$.

3.3 Triadic Loss

Intra-video Discrimination Loss. The output of ϕ at different stages, denoted as ϕ_s, is used to calculate the loss at video level, enhancing the model's ability to learn fine-grained temporal structures. We incorporate a mean squared error smoothing loss, as introduced by [19] and used in [5,27,48]. Considering that actions occurring in an activity video should be temporally contiguous, this loss is applied to the per-frame actions to alleviate over-segmentation. Moreover, we also propose a cyclic variant, based on the assumption (i) described at the beginning of Sect. 3. Specifically, this variant compares the embeddings at the end of the output sequence with those at the start, across different stages of the feature extraction network ϕ. This is driven by the fact that actions often exhibit cyclical patterns in videos. Mathematically, our video-level loss is defined as follows:

$$\mathcal{L}_{video}(i) = \frac{1}{|S||T+1|}\Big(\sum_s \sum_t \left|\log \phi_s(F_i^{t+1}) - \log \phi_s(F_i^{t+1})\right| \\ + \left|\log \phi_s(F_i^{T}) - \log \phi_s(F_i^{0})\right|\Big), \quad (1)$$

where T is the total number of frames and S is the number of stages in ϕ in a video i, $\forall v_i \in V$.

Inter-video Associations Loss. For segment-level learning, we adopt the GTCC loss function proposed by [15], denoted as $\mathcal{L}_{activity}$, to synchronize frames of videos depicting the same activity. We utilize context-adjusted embeddings Ai generated by our context-drop function layer c. Specifically, for each pair $v_i, v_j \in V$ of videos, GTCC computes the probability of dropping v_i^t given v_j for all $t \in T$ using the function c. The loss function is defined as:

$$\mathcal{L}_{GTCC}(v_i|v_j) = \sum_t \Big((1 - \mathrm{P}_{\mathrm{drop}}(v_i^t|A_j)) \cdot \mathcal{L}_{multi-cbr} + \frac{\mathrm{P}_{\mathrm{drop}}(v_i^t|A_j)}{\mathcal{L}_{multi-cbr}}\Big), \quad (2)$$

where $\mathcal{L}_{multi-cbr}$ is a multi-cycle back regression loss, and $\mathrm{P}_{\mathrm{drop}}(v_i^t|A_j)$ is the probability of dropping each video frame v_i^t given A_j (refer to [15] for more details). Our activity loss, $\mathcal{L}_{activity}$ is defined as the sum of $GTCC$ loss of v_i given A_j and vice-versa. This loss leverages the principle of Temporal Cycle Consistency (TCC) [18], ensuring that corresponding frames in videos with identical action sequences are closely aligned in the feature space. This approach addresses variations in action order, redundant actions, and background frames, thereby enhancing the quality of video representations. To the best of our knowledge, this marks the first application of video alignment for temporal action segmentation.

Inter-activity Associations Loss. We learn the global representation of a video clip by using a contrastive loss. We employ contrastive learning to minimize the distance between videos of the same activity while maximizing the distance between videos of different activities. This ensures that videos depicting the same activity are closer in the feature space than videos that are not. The global contrastive loss has the following formulation:

$$\mathcal{L}_{global}(i,j) = (1-y)\cdot d(E_i,E_j) + y\cdot \max(0, m - d(E_i,E_j)) \tag{3}$$

where $d(E_i,E_j)$ denotes the distance between the representations E_i and E_j obtained by ϕ, and $y \in \{0,1\}$ is a binary value such that $y=0$, if the two videos belong to the same activity ($a_i = a_j$), and $y=1$, if they belong to different activities ($a_i \neq a_j$). The margin m ensures sufficient separation between videos of different activities. The term $(1-y)\cdot d(E_i,E_j)$ minimizes the distance for videos of the same activity, while $y\cdot\max(0, m-d(E_i,E_j))$ maximizes the distance for videos of different activities by pushing them apart by at least the margin m.

The combined loss function that governs the training for all pair videos $\{v_i, v_j\} \in V$ of our model is formulated as:

$$\mathcal{L}_{\text{train}}(\phi, c)) = \begin{cases} \alpha\mathcal{L}_{\text{global}}(i,j) + (1-\alpha)\mathcal{L}_{\text{activity}}(i,j) & \\ \qquad\qquad +\beta(\mathcal{L}_{\text{video}}(i) + \mathcal{L}_{\text{video}}(j)), & \text{if } v_i = v_j \\ \mathcal{L}_{\text{global}}(i,j) + \beta(\mathcal{L}_{\text{video}}(i) + \mathcal{L}_{\text{video}}(j)), & \text{if } v_i \neq v_j \end{cases} \tag{4}$$

where α, and β are hyperparameters that balance the contributions of the global, activity, and video loss components. Incorporating this loss in our model allows us to leverage the weak supervision effectively, making the clustering of video frames more discriminative and improving the overall performance of action segmentation and classification tasks in a global manner.

4 Experimental Setup

Datasets. We present results on two well-known datasets used for temporal action segmentation: **Breakfast Action Dataset (BF)** [22] is one of the largest fully annotated collections available for temporal action segmentation.

It includes 1712 videos, featuring 10 activities related to breakfast preparation. These activities are performed by 52 individuals across 18 different kitchens. Each video has an average of 2099 frames. Remarkably, only 7% of the frames are background frames. **Youtube INRIA Instructional Dataset (YTI)** [2] includes 150 instructional videos from YouTube, covering 5 different activities such as changing a car tire, preparing coffee, and performing cardiopulmonary resuscitation (CPR). The videos have an average duration of 2 min. A significant challenge with this dataset is the high proportion of background frames, which make up 63.5% of the total frames.

Features. To ensure a fair comparison with related work, we utilized the same input features as recent methods. For the BF dataset, we used the IDT features [44] provided by the authors of CTE [22] and SCT [34]. These features capture motion information by tracking dense points in the video and computing descriptors such as Histogram of Oriented Gradients, Histogram of Optical Flow (HOF) [25], and Motion Boundary Histogram. Additionally, for further comparison in the BF dataset, we employ I3D features [8] extracted from the Inflated 3D ConvNet, which leverages both spatial and temporal convolutions to learn video representation. For the YTI dataset, we use the same features as [2,14]. These 3000-dimensional feature vectors are formed by concatenating HOF descriptors with features extracted from the VGG16-conv5 layer [39].

Metrics. To evaluate the performance of our temporal action segmentation methods, we employ 1) Mean over Frames (MoF), which calculates the accuracy as the mean percentage of correctly classified frames across all videos, providing a direct indication of overall segmentation performance; 2) F1-Score, which is the harmonic mean of precision and recall, accounting for both false positives and false negatives. Precision is the ratio of correctly predicted action frames to the total predicted action frames, while recall is the ratio of correctly predicted action frames to the total actual action frames; 3) MoF with Background (MoF-BG), which calculates the accuracy considering both action and background frames, essential for understanding how well the segmentation method distinguishes between action and non-action frames, especially given the high proportion of background frames in the YTI dataset. To enable direct comparison, we follow the procedure used in previous work [7,16,23,28,37], reporting results by removing the ratio ($\tau = 75\%$) of the background frames from the video sequence.

Evaluation Setting. In our study, we adopt the global evaluation methodology proposed by [23]. This methodology involves grouping videos into coherent subsets K and representing them using a bag-of-words (BoW) approach. These representations are then clustered into K' groups of pseudo-activities and K subgroups of actions are inferred. Each video is temporally segmented by assigning each frame to one of the ordered groups using the Viterbi decoder. A background model is introduced to deal with irrelevant segments. Throughout the results of this work, the inclusion of BoW and Decoding refers to the integration of the aforementioned global inference process, which we will refer to as the *post-processing protocol*.

For evaluation, we perform a Hungarian matching between the inferred clusters and the ground-truth labels to compute the metrics. Specifically, we assume in the case of the Breakfast dataset $K' = 10$ activity clusters with $K = 5$ subactions per cluster. Subsequently, we match 50 different sub-action clusters with 48 ground-truth sub-action classes, with frames of the leftover clusters set as background. Finally, we assess the accuracy of the unsupervised learning configuration on the YouTube Instructions dataset, employing $K' = 5$ and $K = 9$, subsequently matching 45 distinct sub-action clusters with 47 ground-truth subaction classes.

Training Details. To ensure that each video in our training set has at least one pair from the same activity and one pair from a different activity, we construct the training set by including all possible combinations of videos belonging to the same activity. Since segment-level learning requires a strong initialization to align actions between videos, we adopt a two-stage training approach. Initially, the model is trained with global-level and video-level modules using Eq. 1 and 3, respectively. Subsequently, the model is used to initialize the second stage, where it is trained using the full loss function in Eq. 4. In a stratified fashion, we select a subset of pairs from different activities, ensuring an equal number of same-activity and different-activity pairs. Given a large number of combinations, in each epoch, we take a batch including 50% of the dataset of possible pairs for each epoch. Note that each epoch uses a batch size of 32 pairs for the BF dataset and 8 pairs for the YTI dataset. We simultaneously train a 4-layer feedforward neural network for the drop-context function, c, along with ϕ. To enhance computational efficiency, we down-sample all videos to 256 frames per video by randomly removing frames distributed throughout each video, similar to [24,42]. This technique reduces frame redundancy and ensures that the frames represent the entire video. We use the same parameters as specified in [5,15] for each network. The training process employs the ADAM optimizer, with a learning rate of $2e^{-4}$ and a weight decay of 10^{-4}. For the parameters α and β, we select the values 0.15 and 0.5, respectively.

4.1 Comparative Methods

The method more similar to ours in terms of scope, i.e. global action segmentation, and information used, i.e. activity labels, is CAD [14]. For the sack of completeness, we compute results with a global matching scope of state-of-the-art methods conceived for action segmentation at activity level. These include on the one side unsupervised methods such as ASOT [47], CTE [23] and ASAL [26] that train a network for each activity hence using our same pseudo-labels; on the other side, they include weakly-supervised methods such as ATBA [46] that instead use a transcript for each video, resulting in a much stronger level of supervision.

Table 1. Action Segmentation results on the BF and YTI datasets by applying the Hungarian matching at global-level. The dash indicates "not reported." (*) denotes results computed by ourselves. "F" denotes the type of features used. "D" indicates the use of Viterbi decoding. Both marks denote evaluation as per [23]. The best results are marked in bold.

BF

Supervision	Approach	F	BoW	D	F1	MoF
Unsupervised	CTE [23]	IDT	✓	✓	–	18.5
	ASAL [26]	IDT	✓	✓	–	20.2
	ASOT* [47]	IDT	✓	✓	20.2	21.6
Weak	CAD [14]	IDT	✗	✗	–	10.9
			✓	✗	–	17.7
			✓	✓	–	<u>23.4</u>
	2by2	IDT	✓	✓	**20.6**	**24.6**
Unsupervised	ASOT* [47]	I3D	✓	✓	16.9	18.1
Weak-transcripts	ATBA* [46]	I3D	✓	✓	20.0	17.7
Weak-activity labels	CAD [14]	I3D	✗	✗	–	19.2
	2by2	I3D	✓	✓	**17.5**	**20.7**

YTI

Supervision	**Approach**	**BoW**	**D**	**F1**	**MoF**	**MoF-BG**
Unsupervised	CTE [23]	✓	✓	–	19.4	10.1
	ASOT* [47]	✓	✓	15.26	18.6	9.9
Weak	CAD [14]	✗	✗	12.10	15.7	–
	2by2	✓	✓	**16.53**	**23.6**	**11.4**

5 Results

5.1 Comparison with the State-of-the-Art

Breakfast Dataset (BF). The results obtained by using the IDT features as input demonstrate a consistent performance improvement over prior methods (refer to left-hand Table 1). We achieved a +1.2% improvement in MoF with respect to CAD, highlighting the efficacy of our global training approach with binary labels.

We computed the results at the global level of ASOT [47], by following the evaluation protocol described above. 2by2 outperforms it in terms of MoF by +3% and in terms of F1-score by +0.4%. Similar trends are observed when using I3D features as input. Compared to state-of-the-art methods, the 2by2 framework proves effective regarding MoF and F1-score. ATBA [46] exhibits a higher F1-score but a lower MoF than 2by2, likely due to its use of transcripts for each video, providing stronger supervision with respect to our method but poorer generalization across activities. This could be attributed to the fact that these methods were not specifically designed for global training, highlighting the critical importance of inter-activity learning which is currently lacking in other unsupervised methods.

Inria Instructional Videos (YTI). The performance of our 2by2 framework also shows marked improvements over previous methods on the YTI (refer to right-hand Table 1). We achieve an increase in MoF of +4.2% without background and +1.3% with background. This improvement in the F1 score is likely attributed to the non-repetitive nature of actions within activities in this dataset. Our 2by2 framework effectively enhances segmentation accuracy compared to

Table 2. Ablation studies on the YTI dataset, highlighting the importance of the three loss terms, as well as of the concept of temporal cycles and the initialization with k-means.

YTI

$\mathcal{L}_{\text{video}}$	$\mathcal{L}_{\text{activity}}$	$\mathcal{L}_{\text{global}}$	**MoF**
✓	✓	✓	**23.6**
✗	✓	✓	21.9
✓	✗	✓	22.5
✗	✗	✓	21.8

Base	**23.6**
No k_means init	21.1
No cycled MSE	21.0
No k_means init and cycled	20.4

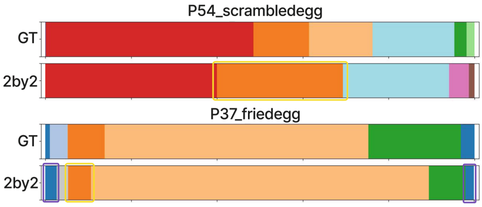


Fig. 3. Examples from BF ("scrambled egg" and "fried egg" activities). Comparison of ground truth (GT) segmentation and our 2by2 framework. 2by2 discovers common action steps across activities (see yellow segments) and captures the cyclic nature of the videos (see purple segments). (Color figure online)

ASOT, the leading unsupervised activity-level segmentation method. Similar to BF, our results underscore the effectiveness of inter-activity training. Furthermore, leveraging global-level training with CAD, we observe significant improvements of +7.9% in MoF and +4.4% in F1 score.

The observed performance improvements in both datasets are likely due to the framework's ability to identify better shared actions among pseudo-activity classes caused by inaccurate pseudo-labels and the enhanced initialization of the Bag of Words (BoW) model through video alignment.

Qualitative Result. In Fig. 3, we observe examples closely aligning with the ground truth segments, accurately capturing both large and small segments. The enhanced segmentation arises from multi-level processing within our framework. The activity-level component (GTCC) facilitates precise segment alignment, while the global aspect improves activity differentiation and reduces misclassification. At the video level, our framework maintains temporal consistency and cyclic patterns, reducing over-segmentation and enhancing alignment.

5.2 Ablation Study

In Table 2, we show the importance of modelling all three levels of learning, by using $\mathcal{L}_{\text{video}}$, $\mathcal{L}_{\text{activity}}$ and $\mathcal{L}_{\text{global}}$. Specifically, we observe that the elimination of the intra-video component significantly impacts our method's performance, highlighting the detrimental effect of relying solely on the global loss. Additionally, since the inter-video component is introduced in the second stage, it becomes clear that robust initialization in the first stage is essential for $\mathcal{L}_{\text{activity}}$

to effectively guide the alignment and segmentation processes. This underscores that the global loss alone in the first stage is insufficient for achieving optimal performance.

Furthermore, we ablate the effect of initializing the activity cluster for the last layer used for $\mathcal{L}_{\text{global}}$ by using k-means instead of random initialization. Additionally, the negative impact of removing the cyclic component from $\mathcal{L}_{\text{video}}$ is evident.

6 Conclusion

This paper introduced 2by2, a novel framework for weakly supervised temporal action segmentation in untrimmed videos encompassing different activities. The proposed architecture consists of a Siamese transformer-based network that takes input pairs of videos and determines if they belong to the same activity or not. If they do, the videos are also temporally aligned. A key innovation of our approach is the direct action alignment between videos, crucial for accurately matching corresponding segments. This is enabled by the Siamese two-stage architecture that ensures robust initialization for temporal alignment. By explicitly modelling intra-video action discrimination, inter-video action associations, and inter-activity action associations, our method significantly outperforms state-of-the-art approaches on the challenging BF and YTI datasets.

Acknowledgements. This work was supported by the grant PRE2020-094714, the project PID2019-110977GA-I00 and the project PID2023-151351NB-I00 funded by MCIN/ AEI /10.13039/501100011033, by "ESF Investing in your future" and by ERDF, UE.

References

1. Aakur, S.N., Sarkar, S.: A perceptual prediction framework for self supervised event segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2019)
2. Alayrac, J.B., Bojanowski, P., Agrawal, N., Sivic, J., Laptev, I., Lacoste-Julien, S.: Unsupervised learning from narrated instruction videos. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 4575–4583 (2016)
3. Andrew, G., Arora, R., Bilmes, J., Livescu, K.: Deep canonical correlation analysis. In: Deep Canonical Correlation Analysis (ICML), pp. 1247–1255 (2013)
4. Arandjelovic, R., Zisserman, A.: Look, listen and learn. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2017)
5. Bahrami, E., Francesca, G., Gall, J.: How much temporal long-term context is needed for action segmentation? In: Proceedings of the IEEE International Conference on Computer Vision (ICCV) (2023)
6. Behrmann, N., Golestaneh, S.A., Kolter, Z., Gall, J., Noroozi, M.: Unified fully and timestamp supervised temporal action segmentation via sequence to sequence translation. In: Proceedings of the European Conference on Computer Vision (ECCV) (2022)

7. Bueno-Benito, E., Tura, B., Dimiccoli, M.: Leveraging triplet loss for unsupervised action segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW) (2023)
8. Carreira, J., Zisserman, A.: Quo vadis, action recognition? A new model and the kinetics dataset. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2017)
9. Chang, X., Tung, F., Mori, G.: Learning discriminative prototypes with dynamic time warping. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2021)
10. Cuturi, M., Blondel, M.: Soft-DTW: a differentiable loss function for time-series. In: International Conference on Machine Learning (ICML) (2017)
11. Dias, C., Dimiccoli, M.: Learning event representations by encoding the temporal context. In: Proceedings of the European Conference on Computer Vision (ECCV) Workshops. vol. 11131, pp. 587–596 (2018)
12. Dimiccoli, M., Wendt, H.: Learning event representations for temporal segmentation of image sequences by dynamic graph embedding. IEEE Trans. Image Process. **30**, 1476–1486 (2020)
13. Ding, G., Sener, F., Yao, A.: Temporal action segmentation: an analysis of modern techniques. IEEE Trans. Pattern Anal. Mach. Intell. **46**(2), 1011–1030 (2023)
14. Ding, G., Yao, A.: Temporal action segmentation with high-level complex activity labels. IEEE Trans. Multimedia **25**, 1928–1939 (2023)
15. Donahue, G., Elhamifar, E.: Learning to predict activity progress by self-supervised video alignment. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2024)
16. Du, Z., Wang, X., Zhou, G., Wang, Q.: Fast and unsupervised action boundary detection for action segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2022)
17. Dvornik, N., Hadji, I., Derpanis, K.G., Garg, A., Jepson, A.D.: Drop-DTW: aligning common signal between sequences while dropping outliers. In: Advances in Neural Information Processing Systems (2021)
18. Dwibedi, D., Aytar, Y., Tompson, J., Sermanet, P., Zisserman, A.: Temporal cycle-consistency learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2019)
19. Farha, Y.A., Gall, J.: MS-TCN: multi-stage temporal convolutional network for action segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2019)
20. Haresh, S., et al.: Learning by aligning videos in time. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2021)
21. He, Y., Yuan, Z., Wu, Y., Cheng, L., Deng, D., Wu, Y.: ViSTec: video modeling for sports technique recognition and tactical analysis. In: Proceedings of the Conference on Artificial Intelligence (AAAI) (2024)
22. Kuehne, H., Arslan, A., Serre, T.: The language of actions: recovering the syntax and semantics of goal-directed human activities. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2014)
23. Kukleva, A., Kuehne, H., Sener, F., Gall, J.: Unsupervised learning of action classes with continuous temporal embedding. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2019)
24. Kumar, S., Haresh, S., Ahmed, A., Konin, A., Zia, M.Z., Tran, Q.H.: Unsupervised action segmentation by joint representation learning and online clustering.

In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2022)

25. Laptev, I., Marszalek, M., Schmid, C., Rozenfeld, B.: Learning realistic human actions from movies. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2008)
26. Li, J., Todorovic, S.: Action shuffle alternating learning for unsupervised action segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2021)
27. Li, S.J., AbuFarha, Y., Liu, Y., Cheng, M.M., Gall, J.: MS-TCN++: multi-stage temporal convolutional network for action segmentation. IEEE Trans. Pattern Anal. Mach. Intell. **45**(6), 6647–6658 (2020)
28. Li, Y., Xue, Z., Xu, H.: OTAS: unsupervised boundary detection for object-centric temporal action segmentation. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV) (2024)
29. Liu, W., Tekin, B., Coskun, H., Vineet, V., Fua, P., Pollefeys, M.: Learning to align sequential actions in the wild. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2022)
30. Lu, Z., Elhamifar, E.: Weakly-supervised action segmentation and alignment via transcript-aware union-of-subspaces learning. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV) (2021)
31. Lu, Z., Elhamifar, E.: Set-supervised action learning in procedural task videos via pairwise order consistency. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 19871–19881 (2022)
32. Lu, Z., Elhamifar, E.: FACT: frame-action cross-attention temporal modeling for efficient action segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2024)
33. Ng, Y.B., Fernando, B.: Weakly supervised action segmentation with effective use of attention and self-attention. Comput. Vis. Image Underst. **213**, 103298 (2021)
34. Rohrbach, M., et al.: Recognizing fine-grained and composite activities using hand-centric features and script data. Int. J. Comput. Vis. (IJCV) **119**, 346–373 (2016)
35. Sarfraz, M.S., Murray, N., Sharma, V., Diba, A., Gool, L.V., Stiefelhagen, R.: Temporally-weighted hierarchical clustering for unsupervised action segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2021)
36. Sarfraz, M.S., Sharma, V., Stiefelhagen, R.: Efficient parameter-free clustering using first neighbor relations. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2019)
37. Sener, F., Yao, A.: Unsupervised learning and segmentation of complex activities from video. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2018)
38. Sermanet, P., et al.: Time-contrastive networks: self-supervised learning from video. In: IEEE International Conference on Robotics and Automation (ICRA) (2018)
39. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)
40. Sirnam Swetha, Hilde Kuehne, Y.S.R., Shah, M.: Unsupervised discriminative embedding for sub-action learning in complex activities. In: IEEE International Conference on Image Processing (ICIP) (2021)
41. Souri, Y., Fayyaz, M., Minciullo, L., Francesca, G., Gall, J.: Fast weakly supervised action segmentation using mutual consistency. IEEE Trans. Pattern Anal. Mach. Intell. **44**(10), 6196–6208 (2022)

42. Tran, Q.H., Mehmood, A., Ahmed, M., Naufil, M., Konin, A., Zia, M.Z.: Permutation-aware activity segmentation via unsupervised frame-to-segment alignment. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV) (2024)
43. VidalMata, R.G., Scheirer, W.J., Kukleva, A., Cox, D.D., Kuehne, H.: Joint visual-temporal embedding for unsupervised learning of actions in untrimmed sequences. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV) (2021)
44. Wang, H., Schmid, C.: Action recognition with improved trajectories. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV) (2013)
45. Wang, Z., et al.: SSCAP: self-supervised co-occurrence action parsing for unsupervised temporal action segmentation. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV) (2022)
46. Xu, A., Zheng, W.S.: Efficient and effective weakly-supervised action segmentation via action-transition-aware boundary alignment. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2024)
47. Xu, M., Gould, S.: Temporally consistent unbalanced optimal transport for unsupervised action segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2024)
48. Yi, F., Wen, H., Jiang, T.: ASFormer: transformer for action segmentation. In: The British Machine Vision Conference (BMVC) (2021)
49. Zhang, R., Wang, S., Duan, Y., Tang, Y., Zhang, Y., Tan, Y.P.: HOI-aware adaptive network for weakly-supervised action segmentation. In: Proceedings of the Conference on Artificial Intelligence (AAAI) (2023)
50. Zuckerman, I., et al.: Depth over RGB: automatic evaluation of open surgery skills using depth camera. Int. J. Comput. Assist. Radiol. Surg. **19**(25), 1–9 (2024)

MDFIDNet: Multi-domain Feature Integration Denoising Network

Debashis Das(✉) and Suman Kumar Maji

Department of Computer Science and Engineering, Indian Institute of Technology,
Patna 801106, Bihar, India
{debashis_2221cs31,smaji}@iitp.ac.in

Abstract. In the realm of computer vision, image denoising remains a formidable challenge with profound implications for fields like medical imaging, remote sensing, and photography. Despite notable advancements in deep learning, there are enduring challenges: current convolutional neural networks (CNNs) frequently struggle with training complexities due to their emphasis on increased network depth. At the same time, these networks often fail to adequately consider the crucial role of gradient information in the denoising process. Furthermore, there is a distinct gap in leveraging transform domain analysis in image denoising. This study addresses these limitations with MDFIDNet, a novel triple-phase attentive fusion network tailored for image denoising. MDFIDNet integrates three independent feature extraction pipelines: a frequency domain processing pipeline (FDP) enhanced by a multi-scale convolutional attention Block (MSCAB), a spatial domain processing pipeline (SDP) focusing on detail feature preservation, and a gradient-domain processing pipeline (GDP) driven by multidirectional gradient information. Experimental validation demonstrates that MDFIDNet surpasses existing benchmarks, exhibiting robust performance across diverse datasets. Comprehensive ablation studies underscore the individual contributions of each network component, elucidating the novel advancements that underpin MDFIDNet's superior denoising efficacy. The source code and further details are available in the https://github.com/debashis15/MDFIDNet.

Keywords: Computer vision · Deep Learning · Image denoising · Gradient information · Real images · Experimentation · Training Complexity

1 Introduction

Image denoising is a highly active area of research in computer vision, focusing on restoring clean images from noisy ones. This process is essential for many real-world applications, as the quality of denoised images profoundly impacts the performance of downstream tasks such as image classification, image segmentation, object detection, and other advanced computer vision tasks [14,24]. Despite its importance, image denoising remains a challenging task due to the

A. Antonacopoulos et al. (Eds.): ICPR 2024, LNCS 15332, pp. 392–407, 2025.
https://doi.org/10.1007/978-3-031-78125-4_27

complexity of real-world scenes and inherent information loss. To address this challenge effectively, our goal is to develop a versatile approach that embodies the following key attributes: 1) end-to-end image denoising; 2) computationally efficient; and 3) applicable to real-world data. In recent years, Several traditional methods leverage geometric features of images, such as sparse coding [7], self-similarity [8], and low-rank estimators [9], to perform denoising. Notably, methods like block matching 3-D filter (BM3D) [5] and weighted nuclear norm minimization (WNNM) [9] are considered state-of-the-art. However, these methods often involve iterative processes, leading to high computational costs and inefficiencies. They also rely on manually crafted priors such as sparsity or NSS, which may not universally apply to all natural images.

With the advent of Convolutional Neural Networks (CNNs), many CNN-based denoising models have effectively addressed the limitations of traditional methods, offering reduced hyperparameters and shorter inference times. The adaptable and robust learning capabilities of CNNs have led to significant advancements in image denoising. Dong *et al.* [6] pioneered the use of CNNs in this domain with SRCNN, which employed three convolutional layers for image super-resolution, greatly improving performance over previous techniques. Similarly, DnCNN [33] was the first CNN-based model to implement batch normalization and residual learning, achieving superior denoising results. Since then, numerous CNN-based denoising methods have been developed. FFDNet [34] effectively handles various noise levels by incorporating a down-sampling operation and a noise level map. CBDNet [10] employs a noise estimation strategy through two sub-networks, enhancing its deep learning approach. DCTNet [11] leverages a DCT transform-based architecture with shrinkage blocks and residual learning to achieve competitive results. Liu et al. [18] introduced a deep multi-level wavelet CNN (MWCNN), which integrates wavelet and U-Net architectures to extract frequency features, and MWDCNN [26] a multi-stage denoiser with wavelet transform further advancing the field of image denoising. Gradient information integration has become instrumental in enhancing denoising methodologies. For instance, Liu et al. developed GradNet [19], a CNN-based framework that combines horizontal and vertical image gradients with DnCNN [33] to effectively preserve essential edge and texture details. In another innovative approach by Li et al. [16], a hybrid denoising model was proposed, leveraging the combination of BM3D [5] and WNNM [9]. This model decomposes noisy images into subbands before applying BM3D [5], achieving robust denoising results.

Despite the impressive learning capabilities of CNNs, early CNN-based denoisers often emphasize uniform feature extraction, which can fail to adequately capture complex image structures and textures, leading to significant performance degradation. To mitigate this limitation, attention mechanisms have been incorporated into network architectures, resulting in promising denoising outcomes. For instance, RIDNet [2], a single-stage denoiser, leverages channel and spatial dependencies within feature maps to enhance performance. Pan *et al.* [22] introduced GrencNet, which utilizes a guided feature domain denoising residual network, dynamic joint attention modules, and an iterative noise correction scheme to effectively address noise in real-world images. NIFBGDNet

[25], another versatile denoiser, employs a dual-path attention-based architecture that uses the negative of the input image as a prior. Similarly, DRANet [29], a dual residual attention network, is designed to handle both synthetic and real-world noise effectively. MPRNet [31] incorporates a multi-stage architecture with encoder-decoder configurations, while MIRNetv2 [32] utilizes a recursive residual design based on multi-scale feature representation. Additionally, APD-Nets [14], a deep encoding-based Regularization Priors (RP) network, achieves superior results through its innovative approach. However, attention-based denoisers frequently encounter challenges due to increased network depth, which leads to a higher number of parameters and extended inference times, making them less suitable for real-world applications.

Recently, vision transformers have gained prominence in visual tasks like denoising due to their ability to capture long-range dependencies via global self-attention mechanisms. IPT [4] employed an encoder-decoder architecture but faced high computational demands. Building on this, SwinIR [17] integrated residual attention with Swin transformer elements, setting new performance benchmarks. Uformer [27], a U-shaped LeWin transformer-based model using multi-scale restoration modulator showed impressive results. However, these transformer-based methods often suffer from substantial computational overhead and increased memory footprint due to their large network architectures.

Motivated by the effective fusion of CNNs and attention mechanisms and driven to address the aforementioned challenges, this paper introduces MDFIDNet (Multi -Domain Feature Integration Denoising Network), a novel approach for advanced image denoising. MDFIDNet leverages three distinct parallel processing phases: the frequency domain processing pipeline (FDP), spatial domain processing pipeline (SDP) and the gradient domain processing pipeline (GDP). FDP employs Discrete Cosine Transform (DCT) and Inverse Discrete Cosine Transform (IDCT) with a multi-scale convolutional Attention Block (MSCAB) to capture transform-domain features. SDP extracts shallow image features, in the spatial domain, crucial for detail preservation. GDP further enhances structural fidelity by leveraging gradient domain information with the help of the gradient sensitive attention block (GSAB). The synergistic integration of neural attentive mechanisms in GSAB significantly reduces noise while preserving essential image details. Experimental results demonstrate that MDFIDNet achieves competitive denoising performance, marking a substantial advancement in image restoration techniques. The primary contributions of the proposed network are outlined as follows:

1. A novel triple-phase feature extraction mechanism operating in the frequency domain, spatial domain and gradient domain.
2. A novel multi-scale convolutional attention block (MSCAB) for extracting features from noisy images, in the frequency domain, across multiple scales.
3. A novel spatial domain processing pipeline (SDP) is proposed to enhance the preservation of spatial structures and finer feature components.
4. Introduction of gradient enhanced neural unit (GENU), which integrates gradient information from input image using horizontal, vertical, primary diago-

Table 1. Notation Table: Describing Operations and Functions

Notation	Description
$\oplus$	Element wise addition operation
$\odot$	Concatenate operation
$\alpha(x)$	LeakyRelu Operation on x-th input
$\circledast(x)$	Conv2D operation on x-th input
$\circledast^{(iD)}(x)$	i-th dilated Conv2D operation on x-th input
$\otimes$	Multiplication operation

nal, and secondary diagonal directions to enrich feature maps. To the best of our knowledge, this investigation into diagonal gradient potential for image denoising is a pioneering effort in the literature.

5. Novel multi stage feature aggregation module (MSFAM) synergestically integrated for enhanced detail preservation at multiple stage Table 1.

2 Problem Formulation and Objective

Image denoising centers around formulating the problem through the degradation model represented in Eq. 1:

$$Y = X + N \tag{1}$$

Here, N denotes the additive white Gaussian noise (AWGN) commonly present in optical image and Y represents the resulting noisy image. The main goal of image denoising is to reconstruct $\hat{X}$, the most accurate approximation of the original image (X), from the observed noisy image (Y) while minimizing distortion. In this context, we approach the problem as a mapping function. Therefore, we present MDFIDNet, a convolutional neural network (CNN)-based approach designed for image denoising. MDFIDNet transforms the image denoising problem into a learning task, focusing on understanding and learning the mapping function that relates the noisy image (Y) to the clean image (X) using extensive training datasets.

3 Proposed Methodology

The architecture of MDFIDNet, depicted in Fig. 1, is tailored for denoising applications, where the input (Y) is a noisy image and the output ($\hat{X}$) is the denoised version. The network is structured into three main phases: the first phase processes the input in the frequency domain to extract features, the second phase operates in the spatial domain, and the third phase focuses on extracting gradient-enriched features from the noisy image. Each phase incorporates multiple attention mechanisms designed to capture specific feature characteristics, as outlined in their respective sections.

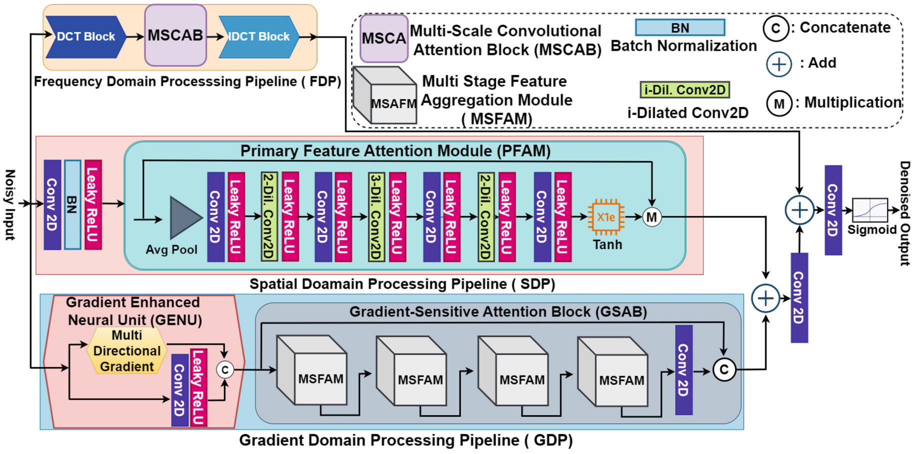

DCT Block
MSCAB
IDCT Block
Frequency Domain Processsing Pipeline (FDP)
MSCA
Multi-Scale Convolutional Attention Block (MSCAB)
BN
Batch Normalization
C: Concatenate
+ : Add
M: Multiplication
MSAFM
Multi Stage Feature Aggregation Module (MSFAM)
i-Dil. Conv2D
i-Dilated Conv2D
Noisy Input
Conv 2D
BN
Leaky ReLU
Primary Feature Attention Module (PFAM)
Avg Pool
Conv 2D
Leaky ReLU
2-Dil. Conv2D
3-Dil. Conv2D
X1e
Tanh
M
Spatial Doamain Processing Pipeline (SDP)
Gradient Enhanced Neural Unit (GENU)
Multi Directional Gradient
C
Gradient-Sensitive Attention Block (GSAB)
MSFAM
Gradient Domain Processing Pipeline (GDP)
Sigmoid
Denoised Output

Fig. 1. The architecture of proposed (MDFIDNet)

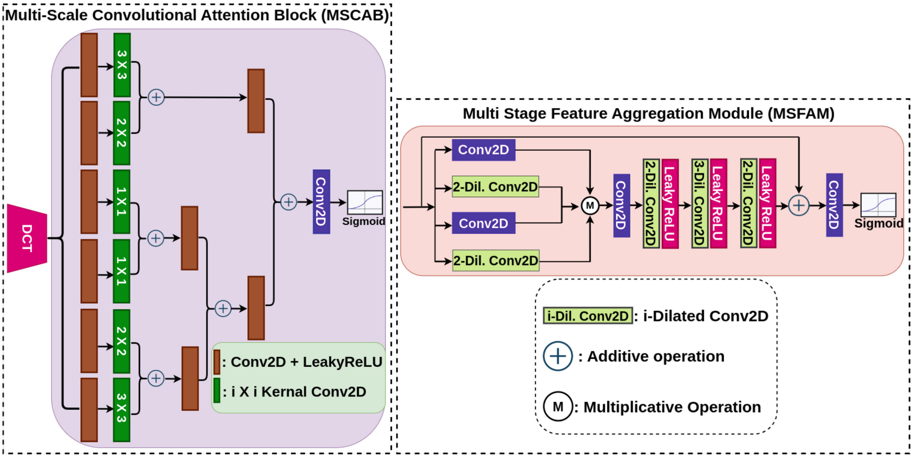

Multi-Scale Convolutional Attention Block (MSCAB)
DCT
3 X 3
2 X 2
1 X 1
Conv2D
Sigmoid
: Conv2D + LeakyReLU
: i X i Kernal Conv2D
Multi Stage Feature Aggregation Module (MSFAM)
Conv2D
2-Dil. Conv2D
M
Leaky ReLU
3-Dil. Conv2D
Sigmoid
i-Dil. Conv2D: i-Dilated Conv2D
: Additive operation
M: Multiplicative Operation

Fig. 2. The architecture of the proposed Multi-Scale Convolutional Attention (MSCAB) and Multi-Stage Feature Aggregation Module (MSFAM).

3.1 Frequency Domain Processing Pipeline (FDP)

The Discrete Cosine Transform (DCT) is used for image denoising due to its ability to convert an image from the spatial domain to the frequency domain, concentrating most of the image's energy into low-frequency components. Noise, typically high-frequency, becomes distinguishable and can be selectively reduced in this domain, preserving the image's structural integrity.

Inspired by DCT's effectiveness, the novel frequency domain processing (FDP) pipeline is introduced, as shown in Fig. 1. The pipeline transforms the input image into the frequency domain using DCT, applies a multi-scale convolutional attention block (MSCAB) for robust feature extraction, and then reverts the image to the spatial domain using the inverse DCT (IDCT). This sequence is mathematically represented as Eq. 2, where $M(\bullet)$ signifies the processing function by the MSCAB block, and $D(\bullet)$ and $D^{-1}(\bullet)$ denote the DCT and IDCT operations, respectively.

$$C(Y) = D^{-1}(M(D(Y))) \tag{2}$$

Multi-scale Convolutional Attention Block (MSCAB). The multi-scale convolutional attention block (MSCAB), shown in Fig. 2, extracts features from input images across multiple scales. The input first passes through a 2D convolution layer followed by Leaky ReLU activation to generate weighted features. These features are then processed through three distinct kernel sizes (3×3, 2×2, and 1×1) to capture intermediate features at various scales, as formulated in Eq. 3.

$$I_i(\bullet) = \circledast_{(i\%3)+1}(\alpha(\circledast(\bullet))) \quad \forall i \in [1, 2, 3, 4, 5, 6] \tag{3}$$

Two of the convolved outputs are combined using element-wise addition to create a comprehensive feature map, which is then processed through another 2D convolution layer followed by Leaky ReLU layer. This intermediate output is added to the remaining convolved output, and the final feature map is obtained after passing through a 2D convolution layer followed by a sigmoid function. This sequence of operations ensures that refined textures remain within the intended range, enhancing denoising performance by preserving critical details, as depicted in Eq. 4, where $\oplus$ denotes element-wise addition and $I_i(\bullet)$ represents the i-th kernel 2D convolution operation.

$$\left\{\begin{array}{l} M(D(Y)) = \sigma(\circledast(\alpha(\circledast(\alpha(\circledast(I_1(D(Y)) \oplus I_2(D(Y))) \oplus \circledast(\alpha(I_3(D(Y)))\oplus \\ I_6(D(Y)))))) \oplus \circledast(\alpha(I_4(D(Y))) \oplus I_5(D(Y))))) \end{array}\right. \tag{4}$$

3.2 Spatial Domain Processing Pipeline (SDP)

The spatial domain denoising pipeline is designed to capture spatial relationships between pixels by extracting relative intra-positional features. The process starts with the noisy input image being processed through an initial combination of 2D convolution, Batch Normalization (β), and Leaky ReLU layers to generate the initial feature map. Next, a primary feature attention module (PFAM) is employed to refine this feature map, ensuring that more useful information is retained. Mathematically, the intra-operation can be expressed as Equation 5, where $P(\bullet))$ represents the processing function of the PFAM block.

$$Q(Y) = P(\alpha(\beta(\circledast(Y)) \tag{5}$$

Primary Feature Attention Module (PFAM). The primary feature attention module (PFAM) uses a progressive refinement mechanism inspired by the hierarchical feature learning seen in biological vision systems, where broad structures are perceived before specific areas are focused on. The module aims to refine pixel attention maps iteratively, moving from coarse to fine details.

The process begins with an average pooling operation that reshapes the input Q(Y) from $C \times H \times W$ to $C \times 1 \times 1$, which can be formulated as Eq. 6.

$$A_p(Q(Y)) = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} Q(Y)_{(i,j)}) \tag{6}$$

This is followed by a 2D convolution layer and Leaky ReLU activation to extract standard features. To reduce the number of parameters and enhance feature diversity, dilated convolutions followed by Leaky ReLU is utilized in alternating stages. Finally, a TanH (τ) activation function, followed by a skip connection via element-wise multiplication, is applied to stabilize the learning, ensure gradient flow, and prevent extreme values. The mathematical formulation of this process is described as Eq. 7.

$$\begin{aligned} P(Q(Y)) = \; & \tau\Big(\alpha\Big(\circledast\Big(\alpha\Big(\circledast^{(2D)}\Big(\alpha\Big(\circledast\Big(\alpha\Big(\circledast^{(3D)}\Big(\alpha\Big(\circledast\Big(\alpha\Big(\circledast^{(2D)}\Big(\alpha\Big(\circledast(\\ & A_p Q(Y)))))))))))))))) \otimes Q(Y) \end{aligned} \tag{7}$$

3.3 Gradient Domain Processing Pipeline (GDP)

The gradient domain processing pipeline is responsible for extracting detailed structural information from the input image using multi-directional gradients. This information is then fed to the subsequent step in a controlled manner, ensuring a precise and nuanced understanding of the image's structure.

Gradient Enhanced Neural Unit (GENU). The gradient enhanced neural unit (GENU) block enhances gradients and detects intensity variations, highlighting textural shifts across different regions within the input image. By analyzing gradient, the GENU block discerns clear image details from noise, with high gradients typically indicating sharp edges. Directional filters (k_h, k_v, k_{d1}, k_{d2}) are utilized for horizontal, vertical, primary, and secondary diagonal directions, respectively, as shown in Eq. 8. These filters emphasize areas with significant gradient magnitudes, thereby enriching the structural representation of the image. The enhanced structural intricacies are then used for further processing within the proposed network.

$$\begin{gathered} G(Y) = \sqrt{[\circledast_{K_h}(Y)]^2 + [\circledast_{K_v}(Y)]^2 + [\circledast_{K_{d1}}(Y)]^2 + [\circledast_{K_{d2}}(Y)]^2} \\ K_h = \begin{bmatrix} -1 & 0 & 1 \\ -2 & 0 & 2 \\ -1 & 0 & 1 \end{bmatrix}, \quad K_v = \begin{bmatrix} -1 & -2 & -1 \\ 0 & 0 & 0 \\ 1 & 2 & 1 \end{bmatrix}, \quad K_{d1} = \begin{bmatrix} 0 & 1 & 2 \\ -1 & 0 & 1 \\ -2 & -1 & 0 \end{bmatrix}, \quad K_{d2} = \begin{bmatrix} -2 & -1 & 0 \\ -1 & 0 & 1 \\ 0 & 1 & 2 \end{bmatrix}. \end{gathered} \tag{8}$$

Following gradient extraction, this module uses a convolutional block for significant feature extraction, consisting of a 2D convolution layer followed by a leaky ReLU activation. The resulting feature map, enriched with convolutional features, is combined with structural gradient information for advanced processing. This enables the denoising algorithm to understand the image's composition and identify specific features for targeted noise reduction. The operational framework is represented in Eq. 9, where $G(Y)$ is the extracted gradient feature.

$$F(Y) = \alpha(\circledast(Y)) \odot G(Y) \tag{9}$$

Gradient Sensitive Attention Block (GSAB). The gradient sensitive attention block integrates processed gradient features using the multi-stage feature aggregation module (MSFAM), which is designed around the principle of cumulative feature integration. Initially, the gradient-enriched feature map is sequentially fused with the MSFAM across four stages to progressively integrate contextual information in a controlled manner. The final feature map undergoes further processing with a single 2D convolutional layer and a skip connection via concatenation to enhance training stability and ensure smooth gradient flow. This integration process is defined by Eq. 10, where $M_i(\bullet)$ denotes the MSFAM processing operation at the respective stage.

$$V(F(Y)) = (M_4(M_3(M_2(M_1(F(Y)))), \odot F(Y)) \tag{10}$$

Multi-stage Feature Aggregation Module (MSFAM). The multi-stage feature aggregation module (MSFAM) is designed to facilitate diverse feature extraction across multiple stages, as illustrated in Fig. 2. Initially, it employs four parallel convolved Layers comprising two 2D convolution and two dilated 2D convolution operations to capture initial features with varying receptive fields which can be represent by Eq. 11.

$$U_1(\bullet) = \circledast(\bullet) \otimes \circledast^{(2D)}(\bullet) \otimes \circledast(\bullet) \otimes \circledast^{(2D)}(\bullet) \tag{11}$$

These operations yield a feature map that integrates diverse feature representations. Following this initial feature capture, the resultant feature map undergoes processing through a single 2D convolution layer to consolidate and refine the extracted features. Subsequently, a sequence of dilated 2D convolution layers, configured as [2, 3, 2], further enhances feature extraction by progressively controlling the receptive fields. To ensure comprehensive feature integration and promote efficient training, the module utilizes skip connections with additive

Table 2. Average PSNR and SSIM for gaussian grayscale dataset

Dataset	σ	BM3D	DNCNN	NIFBGDNet	FFDNet	APD-Net	SWINIR	DRANet	MWDCNN	Proposed
Metric [PSNR/SSIM]										
BSD68	**30**	27.35/0.763	27.17/0.754	28.03/0.783	27.94/0.732	28.10/0.754	27.98/0.760	28.38/0.807	28.32/0.802	**28.71/0.826**
	40	22.44/0.495	25.31/0.682	26.84/0.737	27.55/0.702	27.71/0.733	26.53/0.650	27.54/0.785	27.40/0.781	**27.59/0.801**
	50	18.19/0.316	23.75/0.616	25.99/0.701	26.49/0.677	26.47/0.683	25.83/0.543	**26.85**/0.738	26.68/0.725	**26.85/0.759**
	60	15.73/0.220	22.29/0.546	25.17/0.669	25.24/0.642	25.20/0.629	24.73/0.502	26.43/0.701	26.32/0.687	**26.47/0.703**
Set5	**30**	28.72/0.791	27.84/0.758	30.22/0.849	30.18/0.845	30.26/0.843	29.64/0.820	30.59/0.857	30.23/0.848	**30.64/0.859**
	40	23.12/0.509	25.38/0.683	28.98/0.821	29.03/0.830	29.12/0.828	28.03/0.782	29.66/0.832	29.52/0.825	**29.68/0.833**
	50	18.73/0.324	23.47/0.614	27.98/0.795	28.02/0.791	28.09/0.796	26.64/0.751	28.40/0.803	28.33/0.785	**28.45/0.811**
	60	16.24/0.222	21.85/0.545	26.91/0.756	26.97/0.778	27.01/0.774	24.44/0.703	27.85/0.773	27.77/0.767	**27.91/0.777**
Urban100	**30**	23.34/0.628	26.11/0.620	27.44/0.828	27.59/0.833	27.88/0.831	28.84/0.810	28.89/0.805	27.65/0.749	**30.54/0.838**
	40	22.78/0.562	25.51/0.587	26.76/0.794	27.01/0.798	27.22/0.802	27.08/0.788	27.15/0.771	26.74/0.701	**27.28/0.803**
	50	22.07/0.519	24.77/0.533	25.18/0.712	26.22/0.720	26.73/**0.725**	26.43/0.660	26.58/0.678	26.02/0.650	**26.75**/0.723
	60	21.11/0.498	24.02/0.502	24.59/0.607	25.11/0.634	25.14/0.648	25.92/0.638	25.89/0.640	25.23/0.597	**25.97/0.662**

operations. These connections facilitate the incorporation of global image features learned across multiple stages, thereby enhancing the module's ability to capture intricate image details which is structured as mathematically in Eq. 12.

$$U_2(\bullet) = \alpha(\circledast^{(2D)}(\alpha(\circledast^{(3D)}(\alpha(\circledast^{(2D)}(\circledast(U_1(\bullet))))))) \oplus U_1(\bullet)) \tag{12}$$

The final output of the module is a comprehensive feature map generated by a 2D convolution layer followed by a sigmoid activation function, ensuring normalization of values within the range of 0 to 1. This entire operational framework can be summarized as Eq. 13.

$$U_3(F(Y)) = \sigma(\circledast(U_2(F(Y)))) \tag{13}$$

The final denoised image is achieved by integrating features from the frequency domain processing pipeline (FDP), spatial domain processing pipeline (SDP), and gradient domain processing pipeline (GDP). Initially, the features extracted independently by GDP and SDP undergo convolutional layer processing. This step is iteratively applied to further processed feature of (FDP) block. Finally, a sigmoid activation function yields the ultimate denoised image output.

$$\hat{X} = \sigma(\circledast(\circledast(P(Q(Y)) \oplus V(F(Y))) \oplus C(Y))) \tag{14}$$

4 Loss Function

MDFIDNet trains by selecting patches: I_{clean_i} from pristine images and I_{noisy_i} by adding AWGN to synthetics. For real-world images, which inherently contain noise, patches are extracted and precisely aligned with their corresponding ground truth. The objective is to reconstruct $I_{\text{denoised}_i}{}^* = \text{MDFIDNet}(I_{\text{noisy}_i})$ from the noisy input I_{noisy_i}. The loss function $\mathcal{L}$ is defined as follows:

$$\mathcal{L} \triangleq \frac{1}{2N} \sum_{i=1}^{N} \left\| I^*_{\text{denoised}_i} - I_{\text{clean}_i} \right\|^2 \tag{15}$$

Table 3. Average PSNR and SSIM for gaussian color dataset

Dataset	σ	BM3D	DNCNN	NIFBGDNet	FFDNet	APDNet	SWINIR	DRANet	MWDCNN	Proposed
Metric [PSNR/SSIM]										
CBSD68	**30**	27.21/0.748	28.78/0.853	30.08/0.855	29.78/0.842	30.04/0.853	29.76/0.815	30.08/0.856	29.95/0.851	**30.18/0.860**
	40	26.58/0.738	27.92/0.808	28.71/0.814	28.50/0.803	29.11/0.783	28.50/0.791	28.79/0.820	28.72/0.814	**29.16/0.822**
	50	25.85/0.729	26.49/0.766	27.71/0.779	27.66/0.771	27.88/0.763	27.49/0.769	27.92/**0.782**	27.84/0.773	**27.93**/0.780
	60	24.83/0.695	25.23/0.729	26.90/0.749	26.93/0.747	26.98/0.753	25.37/0.753	27.31/0.772	27.22/0.757	**27.34/0.773**
CUrban100	**30**	23.03/0.615	26.45/0.623	28.14/0.849	28.19/0.856	28.42/0.861	29.53/0.834	29.67/0.866	29.02/0.843	**29.84/0.873**
	40	22.66/0.576	25.01/0.589	27.40/0.782	27.33/0.778	27.51/0.786	28.76/0.801	29.02/0.840	28.69/0.831	**29.22/0.841**
	50	22.03/0.512	24.56/0.521	26.21/0.723	26.56/0.745	26.96/0.756	28.02/0.745	28.52/0.814	28.03/0.801	**28.71/0.820**
	60	21.54/0.489	23.69/0.502	25.88/0.682	26.02/0.667	26.22/0.691	27.62/0.702	**27.86/0.793**	27.75/0.787	27.84/0.790
Manga109	**30**	28.80/0.858	26.78/0.725	31.09/0.897	31.02/0.871	31.03/0.884	29.82/0.868	31.29/0.894	31.22/0.885	**31.30/0.895**
	40	23.54/0.607	23.98/0.638	29.58/0.873	30.21/0.879	30.05/0.871	27.72/0.815	**30.18**/0.874	30.06/0.871	**30.18/0.878**
	50	19.20/0.428	21.82/0.569	28.34/0.849	28.36/0.839	28.45/0.841	26.34/0.793	29.04/0.848	28.95/0.844	**29.07/0.855**
	60	16.63/0.296	20.03/0.493	27.10/0.811	27.03/0.804	27.11/0.801	25.21/0.729	27.82/0.816	27.78/0.807	**27.88/0.818**

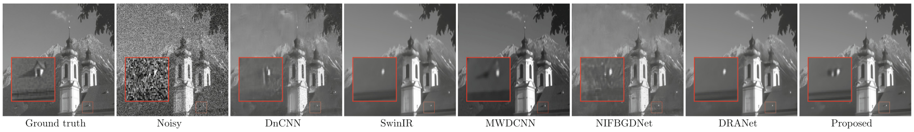


Fig. 3. Visuals of grayscale image denoising on BSD68 dataset($\sigma = 50$).

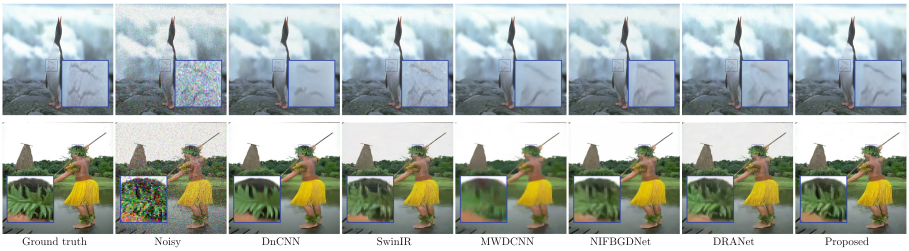
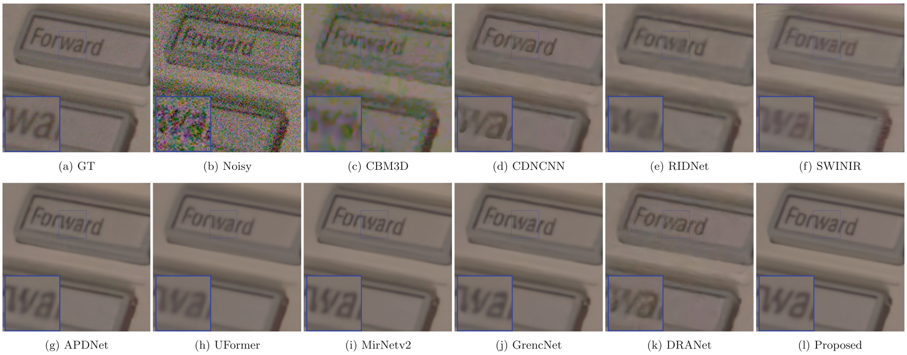


Fig. 4. Visuals of color image denoising on CBSD68 dataset. Top row: Penguine ($\sigma = 30$). Bottom row: Man ($\sigma = 50$).


(a) GT (b) Noisy (c) CBM3D (d) CDNCNN (e) RIDNet (f) SWINIR

(g) APDNet (h) UFormer (i) MirNetv2 (j) GrencNet (k) DRANet (l) Proposed

Fig. 5. Real image denoising results on SIDD validation dataset.

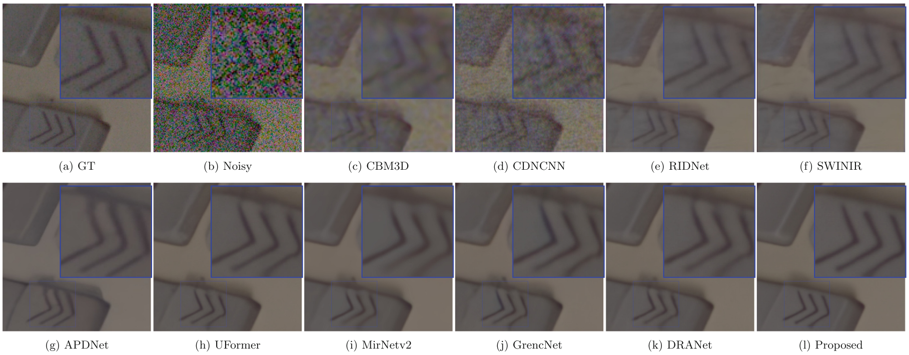


Fig. 6. Real image denoising results on SIDD validation dataset.

5 Experimentation

5.1 Training Setup

The training of the proposed MDFIDNet model was conducted using both synthetic and real datasets. For synthetic noisy image denoising, we utilized BSD400 [20] for grayscale images and CBSD432 [23] for color images, applying Additive White Gaussian Noise (AWGN) with random standard deviations ranging from 0 to 55. The performance of synthetic image denoising was evaluated on five benchmark datasets: Set5 [3], BSD68 [20], CBSD68 [23], and Urban100 [13]. For real image denoising, we employed the SIDD dataset [1], consisting of 512 × 512 image patches, which includes 24,000 training images and 1,280 validation images taken from various smartphone cameras under diverse lighting conditions. The evaluation was performed using the SIDD Validation [1], PolyU [30], and Nam [21], with all patches resized to 256 × 256.

To augment the training data, we applied techniques such as horizontal, vertical flipping and rotations of 90, 180, and 270°, thus enhancing data diversity while retaining essential features. MDFIDNet was trained over 120 epochs with mini-batches of 32 instances, using the Adam optimizer [15] with an initial learning rate of 10^{-3} and a fixed kernel size of 5×5. To ensure stable convergence, the learning rate was dynamically reduced by 0.5 every 25th iteration, and a fixed weight decay was used to prevent overfitting.

5.2 Evaluation on Synthetic Image

The study rigorously evaluated MDFIDNet against several state-of-the-art denoising techniques using PSNR [12] and SSIM [28] metrics across four noise levels ($\sigma = 30, 40, 50, 60$) (Tables 2 and 3). Compared to BM3D [5], DnCNN [33], NIFBGDNet [25], FFDNet [34], APDNet [14], SwinIR [17], DRANet [29], and MWDCNN [26], MDFIDNet consistently achieved the highest PSNR [12] and SSIM [28] scores, indicating superior preservation of original signals and structural details.

Table 4. Average PSNR/SSIM of real images denoising. (Red denotes best)

Dataset	Metric	DnCNN	FFDNet	CBDNet	RIDNet	GrencNet	MPRNet	APD-Nets	MIRNetv2	MCWNNM	DRANet	Uformer	Proposed
SIDD	**PSNR**	29.50	34.22	33.26	38.70	39.42	39.71	39.75	39.82	39.54	39.53	39.89	39.98
	SSIM	0.610	0.855	0.869	0.914	0.957	0.958	0.959	0.959	0.952	0.959	0.960	0.962
PolyU	**PSNR**	36.24	36.84	37.81	38.57	39.69	39.84	39.92	39.85	39.68	39.71	39.85	40.14
	SSIM	0.944	0.892	0.956	0.960	0.965	0.966	0.968	0.967	0.965	0.966	0.968	0.970
Nam	**PSNR**	37.45	37.67	39.09	39.20	39.79	39.97	40.24	40.12	39.72	39.93	40.22	40.28
	SSIM	0.954	0.936	0.969	0.973	0.979	0.981	0.989	0.989	0.986	0.977	0.990	0.992

Table 5. Assessment of computational complexity for various denoising methodologies, using the PolyU testing dataset with dimensions of 512 × 512.

Method	BM3D	MPRNet	FFDNet	Uformer	APDNet	DRANet	GrencNet	MirNet-v2	MWDCNN	MDFIDNet
Device	CPU	GPU	GPU	GPU	GPU	GPU	GPU	GPU	GPU	GPU
Params (M)	–	15.8	**0.87**	51.22	18.61	5.62	5.1	5.9	4.6	4.1
Depth	–	66	64	111	–	48	68	42	36	**28**
MACs	–	587	**71.13**	141.88	212.13	116.36	106.46	106.21	112.21	102.46
FLOPs	–	294	**18.02**	217.56	282.26	187.24	164.76	142.18	174.21	158.33
times (s)	4.23	0.83	**0.28**	0.72	0.80	0.33	0.59	0.48	0.55	0.31
PSNR	36.35	39.84	36.83	39.85	39.92	39.71	39.69	39.85	39.68	**40.14**
SSIM	0.861	0.966	0.892	0.968	0.968	0.966	0.965	0.967	0.965	**0.970**

Table 6. Ablation study on CBSD68 dataset for $\sigma = 30$. (Red denotes best).

Method	w/o frequency Transform	w/o MSCAB	w/o SDP	w/o GENU	w/o GSAB	MDFIDNet
PSNR	30.11	30.08	29.91	29.98	30.01	**30.18**
SSIM	0.854	0.849	0.828	0.818	0.843	**0.860**

Qualitative assessments supported these findings. Visual inspections of grayscale images from the BSD68 dataset at $\sigma = 30$ (Fig. 3) demonstrated MDFIDNet's effectiveness in retaining fine details without artifacts or over-smoothing, unlike its competitors. For color images from (Fig. 4) the CBSD68 dataset at $\sigma = 30$ and 50, MDFIDNet preserved key features better than DnCNN [33], MWDCNN [26], and SwinIR [17], which either oversmoothed or retained noise. Overall, MDFIDNet consistently outperformed other methods both quantitatively and qualitatively, confirming its robustness and efficacy in image denoising tasks.

5.3 Evaluation on Real Image

To evaluate the effectiveness of our method in real-world noise reduction tasks, the proposed method, MDFIDNet, was evaluated on three prominent real-world denoising datasets: SIDD validation [1], PolyU [30], and Nam [21]. Results in Table 4 show that MDFIDNet outperforms existing state-of-the-art techniques in terms of PSNR [12] and SSIM [28] metrics. Visual comparisons in Fig. 5 and Fig. 6 highlight that CBM3D [5], RIDNet [2], and DRANet [29] struggle with maintaining structural details, while SWINiR [17] fails to remove noise. In contrast, our method demonstrates superior structural preservation and visually appealing results. Both quantitative and qualitative assessments confirm that MDFIDNet achieves competitive performance, preserving fine details and structural integrity more effectively than other leading approaches.

5.4 Analysis and Evaluation of Model Complexity

Table 5 compares parameter count, depth, and runtime of denoising methods using PolyU testing images (512 × 512 pixels). MDFIDNet, with a minimal depth of 28 layers, excels in performance metrics like PSNR and SSIM compared to FFDNet [34], known for its efficiency but struggles with complex noise patterns. BM3D [5], on the other hand, requires substantial computational resources due to its block-matching approach. Methods such as MPRNet [31], Uformer [27], and APDNet [14] exhibit significant computation costs owing to their heavy and complex architectures. In contrast, MDFIDNet emerges as a lightweight and effective denoising solution, optimizing performance through a streamlined model design.

5.5 Ablation Study

The ablation study of the proposed network, detailed in Table 6, investigates the impact of including or excluding various blocks within the network architecture. It reveals that the removal of different blocks consistently affects the performance of the proposed method. Notably, the removal of the gradient enhanced neural unit (GENU) significantly decreases the SSIM by ($\downarrow$ 0.042), underscoring the importance of gradient information. Similarly, the removal of the MSCAB and SDP blocks leads to notable performance drops. Overall, the integration of all modules yields the best results, demonstrating the network's peak efficacy.

6 Conclusion

In conclusion, MDFIDNet (Multi-Domain Feature Integration Denoising Network) represents a significant advancement in image denoising within computer vision. By integrating a triple-phase feature extraction approach (frequency domain, spatial domain, and gradient domain) MDFIDNet effectively addresses longstanding challenges in noise reduction. The novel frequency Domain Processing Pipeline (FDP) leverages multi-scale convolutional attention blocks (MSCAB) to extract transform-domain features, enhancing its capability to handle diverse noise patterns. Simultaneously, the spatial domain processing pipeline (SDP) preserves spatial structures and finer details crucial for image fidelity. The introduction of the gradient enhanced neural unit (GENU), which exploits diagonal gradient information, marks a pioneering effort in utilizing gradients for denoising. Experimental results demonstrate that MDFIDNet outperforms existing benchmarks across various datasets, delivering competitive denoising performance with computational efficiency. Further ablation studies confirm the effectiveness of MDFIDNet's design choices, solidifying its role as a leading solution in the realm of image denoising.

References

1. Abdelhamed, A., Lin, S., Brown, M.S.: A high-quality denoising dataset for smartphone cameras. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 1692–1700 (2018). https://doi.org/10.1109/CVPR.2018.00182
2. Anwar, S., Barnes, N.: Real image denoising with feature attention. In: 2019 IEEE/CVF International Conference on Computer Vision (ICCV), pp. 3155–3164 (2019). https://doi.org/10.1109/ICCV.2019.00325
3. Bevilacqua, M., Roumy, A., Guillemot, C.M., Alberi-Morel, M.L.: Low-complexity single-image super-resolution based on nonnegative neighbor embedding. In: British Machine Vision Conference (2012). https://api.semanticscholar.org/CorpusID:5250573
4. Chen, H., et al.: Pre-trained image processing transformer. In: 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE (2021). https://doi.org/10.1109/cvpr46437.2021.01212, http://dx.doi.org/10.1109/CVPR46437.2021.01212
5. Dabov, K., Foi, A., Katkovnik, V., Egiazarian, K.: Image denoising by sparse 3-D transform-domain collaborative filtering. IEEE Trans. Image Process. **16**(8), 2080–2095 (2007). https://doi.org/10.1109/TIP.2007.901238
6. Dong, C., Loy, C.C., He, K., Tang, X.: Image super-resolution using deep convolutional networks (2014)
7. Dong, W., Zhang, L., Shi, G., Li, X.: Nonlocally centralized sparse representation for image restoration. IEEE Trans. Image Process. **22**(4), 1620–1630 (2013). https://doi.org/10.1109/TIP.2012.2235847
8. Fan, L., Li, X., Fan, H., Feng, Y., Zhang, C.: Adaptive texture-preserving denoising method using gradient histogram and nonlocal self-similarity priors. IEEE Trans. Circuits Syst. Video Technol. **29**(11), 3222–3235 (2019). https://doi.org/10.1109/TCSVT.2018.2878794
9. Gu, S., Zhang, L., Zuo, W., Feng, X.: Weighted nuclear norm minimization with application to image denoising. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2862–2869 (2014)
10. Guo, S., Yan, Z., Zhang, K., Zuo, W., Zhang, L.: Toward convolutional blind denoising of real photographs. In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1712–1722 (2019). https://doi.org/10.1109/CVPR.2019.00181
11. Herbreteau, S., Kervrann, C.: Dct2net: an interpretable shallow CNN for image denoising. IEEE Trans. Image Process. **PP**, 1–1 (2021). https://api.semanticscholar.org/CorpusID:236635125
12. Horé, A., Ziou, D.: Image quality metrics: PSNR vs. SSIM. In: 2010 20th International Conference on Pattern Recognition, pp. 2366–2369 (2010). https://doi.org/10.1109/ICPR.2010.579
13. Huang, J.B., Singh, A., Ahuja, N.: Single image super-resolution from transformed self-exemplars. In: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 5197–5206 (2015).https://doi.org/10.1109/CVPR.2015.7299156

14. Jiang, B., Lu, Y., Wang, J., Lu, G., Zhang, D.: Deep image denoising with adaptive priors. IEEE Trans. Circuits Syst. Video Technol. **32**(8), 5124–5136 (2022). https://doi.org/10.1109/TCSVT.2022.3149518
15. Kingma, D., Ba, J.: Adam: A method for stochastic optimization. International Conference on Learning Representations (2014)
16. Li, Z., Liu, H., Cheng, L., Jia, X.: Image denoising algorithm based on gradient domain guided filtering and NSST. IEEE Access **11**, 11923–11933 (2023). https://doi.org/10.1109/ACCESS.2023.3242050
17. Liang, J., Cao, J., Sun, G., Zhang, K., Van Gool, L., Timofte, R.: SwinIR: image restoration using Swin transformer. In: 2021 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW). IEEE (2021). https://doi.org/10.1109/iccvw54120.2021.00210, http://dx.doi.org/10.1109/ICCVW54120.2021.00210
18. Liu, P., Zhang, H., Zhang, K., Lin, L., Zuo, W.: Multi-level wavelet-CNN for image restoration. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops (June 2018)
19. Liu, Y., Anwar, S., Zheng, L., Tian, Q.: GradNet image denoising. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pp. 2140–2149 (2020). https://doi.org/10.1109/CVPRW50498.2020.00262
20. 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 vol.2 (2001). https://doi.org/10.1109/ICCV.2001.937655
21. Nam, S., Hwang, Y., Matsushita, Y., Kim, S.: A holistic approach to cross-channel image noise modeling and its application to image denoising, pp. 1683–1691 (2016). https://doi.org/10.1109/CVPR.2016.186
22. Pan, Y., Ren, C., Wu, X., Huang, J., He, X.: Real image denoising via guided residual estimation and noise correction. IEEE Trans. Circuits Syst. Video Technol. **33**(4), 1994–2000 (2023). https://doi.org/10.1109/TCSVT.2022.3216681
23. Roth, S., Black, M.: Fields of experts. Int. J. Comput. Vis. **82**, 205–229 (2009). https://doi.org/10.1007/s11263-008-0197-6
24. Thakur, R.K., Maji, S.K.: Blind gaussian deep denoiser network using multi-scale pixel attention. In: 2022 IEEE International Conference on Visual Communications and Image Processing (VCIP), pp. 1–5 (2022). https://doi.org/10.1109/VCIP56404.2022.10008856
25. Thakur, R.K., Maji, S.K.: Multi scale pixel attention and feature extraction based neural network for image denoising. Pattern Recogn. **141**, 109603 (2023)
26. Tian, C., Zheng, M., Zuo, W., Zhang, B., Zhang, Y., Zhang, D.: Multi-stage image denoising with the wavelet transform. Pattern Recogn. **134**, 109050 (2023)
27. 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 (CVPR), pp. 17683–17693 (2022)
28. Wang, Z., Bovik, A., Sheikh, H., Simoncelli, E.: Image quality assessment: from error visibility to structural similarity. IEEE Trans. Image Process. **13**, 600–612 (2004). https://doi.org/10.1109/TIP.2003.819861
29. Wu, W., Liu, S., Xia, Y., Zhang, Y.: Dual residual attention network for image denoising. Pattern Recogn. **149**, 110291 (2024)
30. Xu, J., Li, H., Liang, Z., Zhang, D., Zhang, L.: Real-world noisy image denoising: A new benchmark (2018)

31. Zamir, S.W., et al.: Multi-stage progressive image restoration (2021)
32. Zamir, S.W., et al.: Learning enriched features for fast image restoration and enhancement. IEEE Trans. Pattern Anal. Mach. Intell. **45**, 1934–1948 (TPAMI) (2022)
33. Zhang, K., Zuo, W., Chen, Y., Meng, D., Zhang, L.: Beyond a gaussian denoiser: residual learning of deep CNN for image denoising. IEEE Trans. Image Process. **26**(7), 3142–3155 (2017)
34. Zhang, K., Zuo, W., Zhang, L.: FFDNET: toward a fast and flexible solution for CNN-based image denoising. IEEE Trans. Image Process. **27**(9), 4608–4622 (2018)

Dynamic Resolution Guidance for Facial Expression Recognition

Songpan Wang(✉) and Xu Li

School of Information and Software Engineering,University of Electronic Science and Technology of China, Chengdu 610054, Sichuan, China
965728310@qq.com

Abstract. Facial expression recognition (FER) plays a crucial role in human-computer interaction and emotion analysis. However, recognizing expressions in low-resolution images remains a significant challenge. This paper introduces a practical method called Dynamic Resolution Guidance for Facial Expression Recognition (DRGFER) to effectively recognize facial expressions in images with varying resolutions without compromising the accuracy of the FER model. Our framework comprises two main components: the Resolution Recognition Network (RRN) and the Multi-Resolution Adaptation Facial Expression Recognition Network (MRAFER). The RRN determines the resolution of the input image, and the MRAFER assigns the image to the most suitable facial expression recognition network according to its resolution. We evaluated the performance of DRGFER on two widely used datasets, RAF-DB and FERPlus. The results demonstrate that our method maintains optimal model performance at each resolution and outperforms alternative resolution-handling approaches. The proposed framework exhibits robustness against variations in both resolution and facial expressions, offering a promising solution for real-world applications.

Keywords: Facial Expression Recognition · Dynamic Resolution Guidance · Resolution Recognition Network

1 Introduction

Facial expression recognition (FER) is an essential task in video analysis and image understanding, with widespread applications in various fields [6,13,19]. In recent years, FER methods have evolved by employing Convolutional Neural Network (CNN)-based backbone networks to achieve robust feature extraction and facial expression classification is typically conducted using fully connected layers, Support Vector Machines (SVMs), and other similar approaches. Notably, networks such as ResNet [5], Inception network [18], and others have demonstrated impressive feature extraction capabilities, leading to satisfactory performance in training models with individual and static facial images as input.

S. Wang and X. Li—contribute equally to this work.

A. Antonacopoulos et al. (Eds.): ICPR 2024, LNCS 15332, pp. 408–422, 2025.
https://doi.org/10.1007/978-3-031-78125-4_28

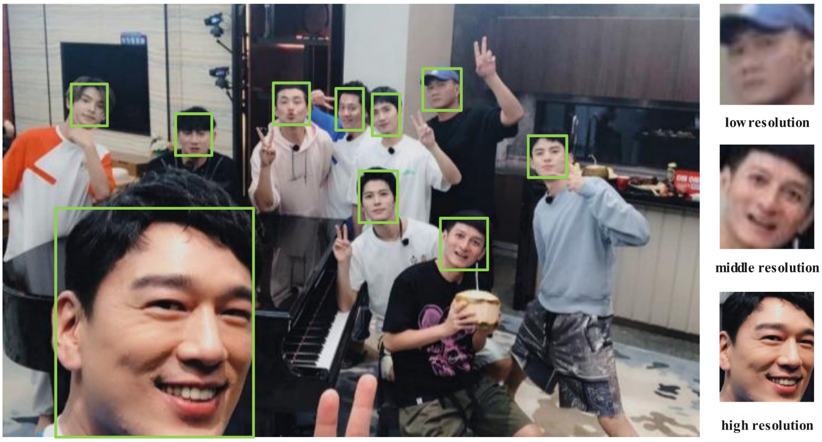


Fig. 1. This is a group photo featuring Chinese celebrities. Due to the shooting angle and distance, the resolution of each individual's face varies. We have selected facial images of three celebrities on the right for an intuitive visual comparison. The images demonstrate the differences in clarity at three distinct resolutions: high, medium, and low.

However, real-world crowd scenes present numerous challenges for FER. One primary challenge is the prevalence of low-resolution images, which can cause the loss of vital feature information, leading to decreased discrimination capabilities. Additionally, as the resolution declines, the feature distribution changes, posing another hurdle for FER in crowd scenes. Specifically, in crowd scenes, facial images of different individuals vary in size (as shown in Fig. 1), presenting a significant challenge in achieving high performance with a single FER model. The reduction in image resolution can be traced to limitations in camera equipment quality and the distance between the subject and the lens. As a result, captured facial images display varying sizes. Figure 5 presents two examples of facial expressions at various resolutions. While the overall expression remains discernible at lower resolutions, the emotional information's characteristics differ significantly from those at higher resolutions.

Image super-resolution (ISR) technology can recover high-resolution images with abundant details from low-resolution images, as demonstrated in previous studies [4,7,9,11,22,23]. In some instances, ISR methods have been applied to enhance low-resolution images to improve performance in FER tasks [3,12]. However, earlier studies [8] have focused mainly on improving the accuracy of the model at a fixed resolution, which can restrict the adaptability of the model to data with varying resolutions. However, the real-world application of low-resolution facial expression recognition algorithms has received insufficient attention, and few studies have focused on applying expression classification models at varying resolutions.

Due to the inherent characteristics of convolutional neural networks, it is challenging to apply them to a wide range of data with different resolutions simultaneously. As a result, using a model trained on a specific resolution or one that has been adapted to incorporate varying resolutions directly is unlikely to yield optimal performance. To address the aforementioned challenges, this paper initially investigates adaptation algorithms at varying resolutions and confirms that it is quite difficult to employ a single model for handling facial expression recognition problems across different resolutions. Subsequently, we propose the **D**ynamic **R**esolution **G**uidance for F**a**cial **E**xpression **R**ecognition (**DRGFER**) framework that can automatically identify the resolution of the input facial image and forward it to the corresponding FER network for recognition. To determine the resolution of each face, a **R**esolution **R**ecognition **N**etwork (**RRN**) is introduced, and **M**ulti-**R**esolution **A**daptation **F**acial **E**xpression **R**ecognition Network (**MRAFER**) will classify expressions based on resolution decisions. Finally, we validate our proposed framework using several widely adopted facial expression datasets, and the experimental results show that our algorithm achieves superior performance.

2 Related Work

Facial expression recognition (FER) [2] has become an important issue with extensive applications in various tasks. However, recognizing facial expressions in low-resolution images poses significant challenges, particularly under realistic conditions where environmental factors and image capture equipment affect image quality. Current FER networks primarily focus on ideal-resolution images, leading to decreased recognition accuracy for low-resolution images, which are common in practical scenarios like surveillance camera footage.

To address this limitation, Jie Shao et al. [17] introduced an edge-aware feedback convolutional neural network (E-FCNN) for recognizing facial expressions in low-resolution images. The E-FCNN incorporates feedback connections between convolutional layers and employs edge-aware convolutional layers to capture detailed information. Wu Gang et al. [20] investigated sample construction and feature embedding, proposing a task-friendly embedding network based on adversarial learning. This network facilitated better reconstruction of lost high-frequency information by generating information-rich positive samples and challenging negative samples in the frequency space. This approach enhanced the model's adaptability to basic-level tasks requiring rich texture and contextual information, thereby advancing research in single-image super-resolution (SISR). However, the single-image super-resolution model struggles to effectively handle multi-scale images, with its reconstruction performance being significantly influenced by the reduced resolution of the input image. Consequently, it is unable to ensure the discriminative sufficiency of recovered features for specific tasks such as object detection and expression classification. Nan Fang et al. [14] also proposed a feature super-resolution-based FER method and employed a novel GAN training strategy that directed the model's attention toward samples that were difficult to classify into the corresponding categories. However, they only focus solely on fixed-resolution problems without considering real-world applications and the implementation of models that adapt to input images with varying resolutions.

The recognition of facial expressions in multi scale low-resolution images has been largely overlooked. Previous approaches treated this as a separate task, training a distinct model for each resolution, which is inefficient. These methods ignore the fact that in real-world scenarios, the resolution of the acquired image data is unknown, making it impossible to determine which model should process the current image. Moreover, existing super-resolution methods require the input image resolution as a priori, which is impractical in the application stage. To address these limitations, we propose the DRGFER framework, which automatically identifies the resolution of the input facial image and forwards it to the corresponding FER network for recognition. So, our method is orthogonal to existing super-resolution methods and addresses different problems.

3 Single Model Adaptation

We explored various methods to enable a single FER to effectively adapt to multi low-resolution facial expression images.

Multi Scale Training (MSTrain). This approach represents a straightforward and essential methodology [15]. By incorporating data augmentation techniques into the training process, a diverse range of low-resolution facial expression image data can be effectively simulated. The underlying objective is to enable the neural network to effectively adapt to these varying resolutions, thereby facilitating targeted training specifically tailored to a particular resolution setting. Regrettably, despite its initial promise, this method did not yield the desired outcome. In fact, it resulted in a noticeable decrease in the model's accuracy across different resolutions.

Domain Adaptive. Domain adaptation [21] has emerged as a prominent research area in recent years. It primarily addresses the effects of data distribution discrepancies on the performance of machine learning models. This concept can be applied to the challenge of multi-scale, low-resolution facial expression recognition. Although data resolutions may vary, leading to distribution biases, the representations used for classification exhibit similarities. Domain adaptation primarily employs feature vectors to accomplish two distinct recognition tasks: the original task of expression recognition and domain recognition, where different resolutions represent separate domains. The training process is based on adversarial learning, to enable the feature encoder to deceive the domain recognizer. The aim is to enable the domain identifier to treat images of different resolutions as equivalent, allowing expression features from varying resolutions to be mapped onto a shared feature space, thereby enabling the classifier to recognize. But, we discovered that this approach cannot prevent a decline in accuracy.

Resolution-Aware BN. Zhu et al. [24] conducted research and determined that varying resolution data exhibit distribution shifts, making it challenging for a single neural network model to adapt to multiple resolutions concurrently. To address this issue, they introduced multiple independent BatchNorm modules in parallel following the convolutional layer, as opposed to the traditional approach of using a single convolutional layer and BatchNorm module. For images of differing resolutions, each respective BatchNorm layer performs an independent normalization operation on the current data. This technique enables the authors to project data from various resolutions into a consistent latent space, thereby reducing the distribution discrepancy between different resolutions. We applied this method to facial expression recognition as well. However, our experimentation showed that this approach does not alleviate the issue of a single model's recognition accuracy degradation when applied to multiple-resolution scenarios.

Upon investigating several aforementioned techniques, we discovered that certain methods, although showing advantages in some studies, do not yield satisfactory results in practical applications in the multi-scale low-resolution facial expression recognition scenario.

4 Methodology

Our proposed method is both simple and practical, Dynamic Resolution Guidance for Facial Expression Recognition (DRGFER). By utilizing our framework, it is possible to achieve end-to-end automatic recognition of multi scale low-resolution facial expressions without compromising the accuracy of the FER model. As depicted in Fig. 2, our proposed framework comprises two stages. Initially, the Resolution Recognition Network(RRN) is utilized to determine the resolution of the input facial expression image. Subsequently, a binarization operation is performed to convert the network's output into a 0–1 vector. The original image, denoted as I_i, and the binarized vector, denoted as $\mathbf{r}_i$, are then fed into the Multi-Resolution Adaptation Facial Expression Recognition Network(MRAFER) as a pair. The network model will automatically select the appropriate facial expression recognition network according to $\mathbf{r}_i$ and ultimately generate the recognition result.

4.1 Resolution Recognition Network

To address the problem of facial expression recognition at different resolutions, we first propose the RRN to guide the subsequent recognition model for more accurate classification of facial expressions.

RRN Architecture. We employ the ResNet18, which uses the same architecture as the FER network used later in the process. The structure of the network, as shown in Fig. 3a, is divided into six parts: stage 1, stages 2–5, and the final resolution predictor. The first stage consists of a convolutional layer, batch normalization layer, ReLU, and max-pooling layer, which extract low-level features

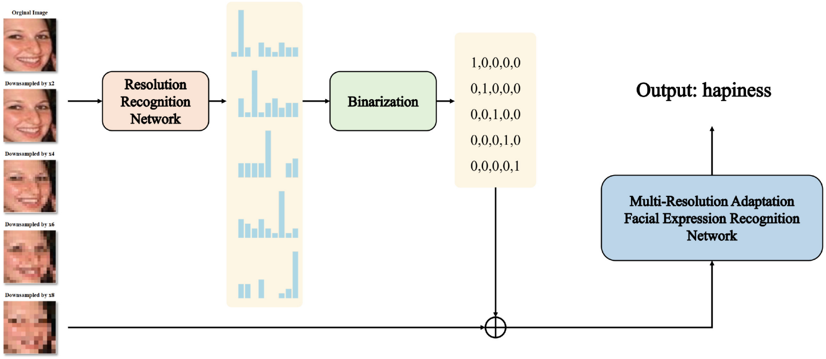


Fig. 2. This is pipeline of our proposed method.

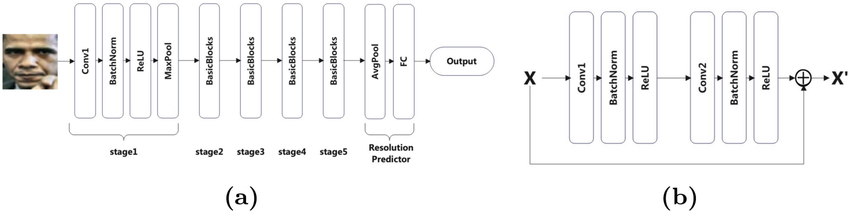


Fig. 3. (a) The structure of RRN, which is based on ResNet18, (b) The Detail of BasicBlock.

from the image while performing the downsampling operation twice. As a result, after the first stage of computation, the feature map resolution is only 1/4 of the input image. The subsequent four stages are composed of BasicBlocks, with two BasicBlocks in each stage. Each BasicBlock is a residual module, as illustrated in Fig. 3a, consisting of two convolution operations, after each convolution operation, BatchNorm is employed to normalize the data, followed by the ReLU operation, which performs nonlinear mapping on the data. The most crucial aspect is that the input and output of the module are added together, utilizing the residual concept to guide the weights in the module during training. Among stages 2–5, only stage 2 does not include any downsampling operations, while the others do. Finally, the resolution predictor consists of an average pooling layer and a fully connected layer. The feature map is converted into a feature vector through average pooling, followed by a predictor to determine the resolution of each image.

The model's output is an unnormalized vector for classification purposes, which is a kind of distribution, we can donate it as $\mathbf{p}^i$ for I_i image.

Loss. Essentially, our RRN is a classification task. Therefore, we employ Softmax to normalize the output vector and utilize the cross-entropy loss function to guide the learning process for this specific component, by Eq.(1) and Eq.(2).

$$\hat{y}_j^i = \frac{\exp(p_j^i)}{\sum_{k=1}^{C} \exp(p_k^i)} \tag{1}$$

$$L_{RRN}(\mathbf{y}^i, \hat{\mathbf{y}}^i) = -\sum_{j=1}^{C} y_j^i \log(\hat{y}_j^i) \tag{2}$$

where C represents the total number of distinct resolution categories, $\hat{y}_j^i$ denotes the probability of the i^{th} image belonging to the j^{th} resolution category, $\hat{\mathbf{y}}^i = \{\hat{y}_0^i, \hat{y}_1^i, ..., \hat{y}_C^i\}$ represents the vector of probabilities for the i^{th} image across all resolution categories, and $\mathbf{y}^i$ is the ground truth.

Binarization. The binarization operation is employed to convert the vector output by the network, as the vector output by the RRN cannot be directly used by our MRAFER. This binarization operation does not rely on a preset threshold, instead, it is based on the maximum value. In this approach, the element with the maximum value in the vector is set to 1, while all other elements are set to 0, the whole process can be defined as the following:

$$\mathbf{r}_i = \begin{cases} 1 & \text{if } j = \mathbf{argmax}(\hat{\mathbf{y}}^i) \\ 0 & \text{otherwise} \end{cases} \tag{3}$$

In this equation, $\mathbf{r}_i$ represents the resolution desicion for image I_i. The operation sets the element j to 1 if it is equal to the maximum value in the corresponding $\hat{\mathbf{y}}^i$, and to 0 otherwise. This results in a binary vector, which can be effectively utilized by the MRAFER to select the appropriate FER_Block for facial expression recognition.

4.2 Multi-resolution Adaptation Facial Expression Recognition Network

As illustrated in Fig. 4, our Multi-Resolution Adaptation Facial Expression Recognition(MRAFER) is comprised of three main components: Assign, FER_Block, and Gather. First, our Assign module traverses the resolution predictions in the entire batch data, then combines images with different resolutions into new batch data, and sends different batches to the corresponding FER_Block. The network structure of our FER module is shown in Fig. 3b. The structure is the same as that of RRN. The only difference is that the number of outputs of the last fully connected layer is different. This is related to the dataset used. Finally, we need to splice the batch data predicted by each FER_Block and use the Gather operation. We divide the whole process into the following steps:

1) Traverse the resolution predictions $\mathbf{r}_i$, in the entire batch data, B.
2) Grouping images with different resolutions into a new batch data, $B' = \{B_1, B_2, ..., B_k\}$, k is the number of FER_Blocks.
3) Send B' to the corresponding FER_Blocks.

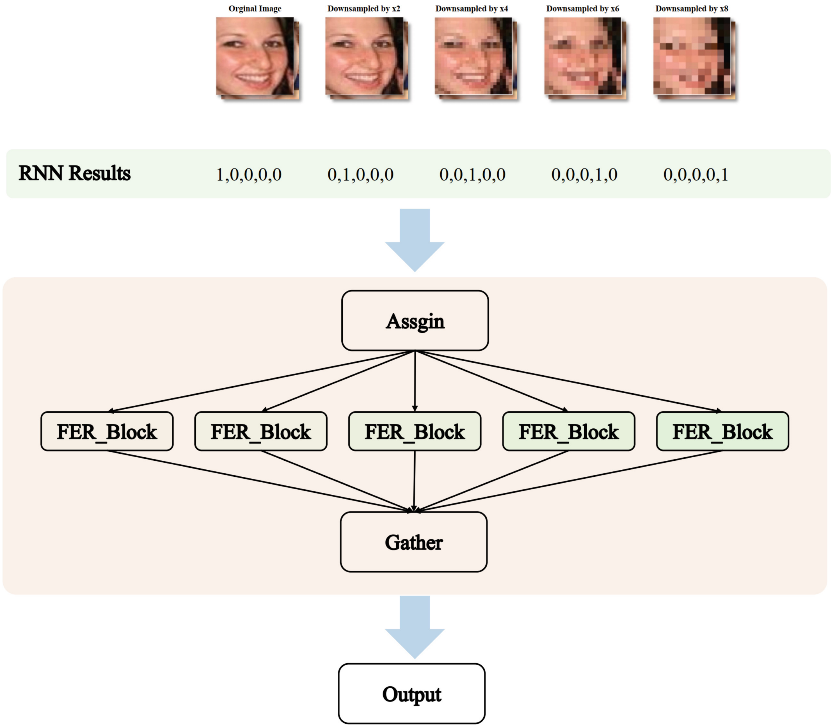


Fig. 4. Multi-Resolution Adaptation Facial Expression Recognition Network.

4) Process the images through the FER_Block.
5) Obtain the predictions B_k^p for each B_k and splice the batch data predicted by each FER_Block into a single output B^p.

$$B' = \mathbf{Assgin}(B) \tag{4}$$

$$B_k = \{I_i | \mathbf{r}_i[k] = 1\} \tag{5}$$

$$B^p = \mathbf{Gather}(\{B_1^p, B_2^p, ..., B_k^p\}) \tag{6}$$

$$B^p[\mathbf{r}_i[k]] = B_k^p \tag{7}$$

We can use Eq.(4) and Eq.(7) to define two operations, Assign and Gather. Equation(5) provides the details of Eq.(4), B_k represents the set of images I_i with their binarized vector $\mathbf{r}_i[k]$ equal to 1. This indicates that they belong to the k-th resolution group. And Eq.(7) provides the details of Eq.(6), $B^p[\mathbf{r}_i[k]]$ denotes the prediction for each image I_i in the k-th resolution group. The Gather operation assigns the prediction from the corresponding resolution group B_k^p to the final output B^p.

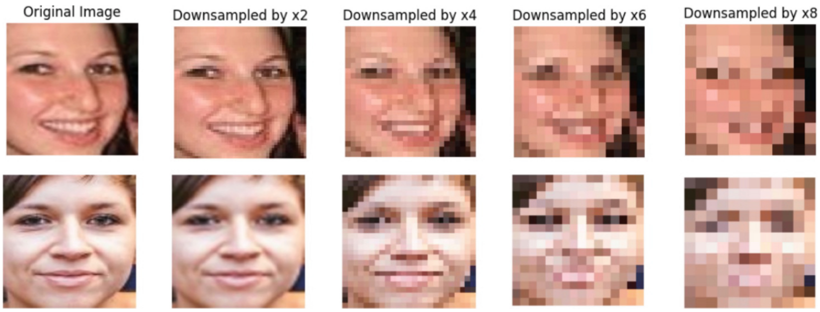


Fig. 5. Two samples from the RAF-DB dataset are visualized. The first column presents the original size image, while the subsequent columns display downsampling to the corresponding magnification.

5 Experiment

5.1 Dataset

To assess the performance of expression recognition, we utilize the RAF-DB [10] and FERPlus [1] datasets in our experiments. **RAF-DB** was compiled using various search engines, and approximately 40 annotators independently labeled each image. The dataset comprises 15,339 images labeled with seven basic emotion categories, with 12,271 designated for training and 3,068 for validation. **FERPlus** is an extension of FER2013, as used in the ICML 2013 Challenges. It is a large-scale dataset collected via the Google search engine, containing 28,709 training images, 3,589 validation images, and 3,589 test images, each resized to 48×48 pixels. The dataset includes an additional class, contempt, resulting in a total of 8 classes. The overall sample accuracy serves as the performance metric.

Similar to most super-resolution studies, we apply a bicubic kernel function to downsample high-resolution images and obtain low-resolution counterparts. The original input size is 100×100 pixels, and we achieve low-resolution images by employing integer down-sample factors of $\times 2$, $\times 4$, $\times 6$, and $\times 8$. Consequently, the total pixel count is reduced to 1/4, 1/16, 1/36, and 1/64, Fig. 5 displays several examples.

5.2 Experiment Setup

Baselines. To evaluate the performance of DRGFER, we compare it with simple max-pooling (Max), average-pooling (Mean) strategies, and the previously mentioned multi-scale augmentation training (MSTrain), domain adaptation (DA), and resolution-aware batch normalization (RA-BN) methods. In our experiments, we use accuracy to evaluate different methods.

For the Max method, we perform a max pooling operation on the logits output by the FER networks trained on different resolutions. This approach selects the most confident prediction among the networks, assuming that the

Table 1. Results of accuracy on RAF-DB with different resolution.

Accuracy Methods Ratio	Mean	Max	RA-BN	DA	MSTrain	DRGFER
×1	86.11%	86.64%	86.96%	81.10%	85.88%	**89.24**%
×2	86.34%	86.28%	86.73%	81.10%	85.91%	**88.23**%
×4	83.93%	82.92%	84.41%	77.22%	84.35%	**85.30**%
×6	75.98%	75.62%	80.18%	69.04%	80.93%	**81.91**%
×8	66.85%	70.47%	76.43%	61.01%	77.18%	**77.35**%
Mean	79.84%	80.38%	82.94%	73.89%	82.85 %	**84.41**%

network trained on the closest resolution to the input image will provide the most accurate prediction. Similarly, for the Mean method, we average the logits output by the FER networks trained on different resolutions. This approach gives equal weight to the predictions from all networks, assuming that the collective knowledge from various resolutions can contribute to a more robust prediction. Both Max and Mean methods aim to leverage the information from multiple resolution-specific networks to improve the overall facial expression recognition performance.

Implementation Details. We resize the input images to 224 pixels, and use random horizontal flipping for data argumentation. We implement the code using the PyTorch [16] framework. All experiments are conducted with a batch size of 256, a learning rate of 3e-4, and the Adam optimizer for 80 epochs. Our experiments are performed on an Nvidia 1080Ti GPU.

5.3 Facial Expression Recognition Result

The Table 1 presents the total accuracy results of six different methods (Mean, Max, RA-BN, DA, MSTrain, and DRGFER) on the RAF-DB dataset, with varying input image resolutions represented by down-sampling ratios (×1, ×2, ×4, ×6, and ×8). The best results in each row are highlighted in bold.

For the highest resolution (×1 down-sampling ratio), DRGFER achieves an impressive accuracy of 89.24%, surpassing the second-best method, RA-BN, by a notable margin of 2.28%. This suggests that DRGFER is capable of extracting fine-grained features and making accurate predictions when provided with high-quality input images. As the down-sampling ratio increases and the image resolution decreases, the performance of all methods declines. However, DRGFER maintains its superior performance, with accuracies of 88.23%, 85.30%, 81.91%, and 77.35% for ×2, ×4, ×6, and ×8 down-sampling ratios, respectively. The consistent lead of DRGFER over other methods across all resolutions highlights its ability to effectively handle the challenges posed by low-resolution images.

In the last row, the table displays the mean accuracy for each method across all down-sampling ratios. DRGFER achieves the highest mean accu-

Table 2. Results of accuracy on FERPlus with different resolutions.

Accuracy Methods / Ratio	Mean	Max	RA-BN	DA	MSTrain	DRGFER
x1	82.66%	80.86%	83.85%	80.77%	82.42%	**84.12**%
x2	83.18%	81.24%	83.27%	80.37%	82.54%	**83.76**%
x4	80.05%	79.53%	82.02%	71.90%	81.47%	**82.23**%
x6	70.94%	73.41%	78.48%	49.26%	77.50%	**78.48**%
x8	62.01%	68.04%	74.05%	40.91%	75.01%	**75.01**%
Mean	75.77%	76.61%	80.33%	64.64%	79.79 %	**80.72**%

racy of 84.41%, followed by MSTrain at 82.85%, RA-BN at 82.94%, Max at 80.38%, Mean at 79.84%, and DA at 73.89%. The results indicate that the DRGFER method consistently outperforms the other tested approaches across various input image resolutions. From Table 1, the experimental results provide strong evidence for the effectiveness of the proposed DRGFER method in facial expression recognition tasks, particularly in scenarios involving varying image resolutions. The ability of our DRGFER to maintain high accuracy across different down-sampling ratios and its notable performance lead over other methods highlight its potential for real-world applications where image quality may vary significantly.

Table 2 presents experiment results conducted on the FERPlus dataset, while all other settings remain the same as in Table 1.

At the original resolution (×1), DRGFER obtains an accuracy of 84.12%, outperforming the second-best method, RA-BN, by a small margin of 0.27%. This indicates that DRGFER is capable of effectively capturing and utilizing the fine-grained details present in high-resolution images for accurate facial expression recognition. As the down-sampling ratio increases, the performance of all methods generally declines due to the loss of image quality and information same as in RAF-DB dataset. However, DRGFER maintains its competitive edge, securing the top position in most cases. For instance, at the ×4 down-sampling ratio, DRGFER achieves an accuracy of 82.23%, surpassing the second-best method, RA-BN, by 0.21%. This highlights the robustness of DRGFER in handling moderately degraded image resolutions. Interestingly, at the ×6 down-sampling ratio, DRGFER and RA-BN obtain the same accuracy of 78.48%, outperforming other methods by a significant margin. This suggests that our DRGFER is particularly effective in extracting meaningful features from low-resolution images, enabling accurate facial expression recognition even in challenging scenarios. At the lowest resolution (×8), DRGFER and MSTrain achieve the highest accuracy of 75.01%, demonstrating their ability to maintain a relatively high performance even when the image quality is drastically reduced. This is particularly impressive considering the substantial performance drop experienced by other methods, such as DA, which obtains an accuracy of only 40.91%.

The mean accuracy across all resolutions further confirms the overall superiority of DRGFER, with an average accuracy of 80.72%, indicates that DRGFER not only excels at specific resolutions but also maintains a consistently high level of performance across a wide range of image resolutions. the experimental results on the FERPlus dataset validate the effectiveness and robustness of the proposed DRGFER method for facial expression recognition under varying image resolutions. DRGFER's ability to achieve competitive performance across different down-sampling ratios, particularly in challenging low-resolution scenarios, highlights its potential for real-world applications where image quality may be compromised. The consistent performance of DRGFER across both RAF-DB and FERPlus datasets demonstrates its generalizability and adaptability to different data distributions and characteristics.

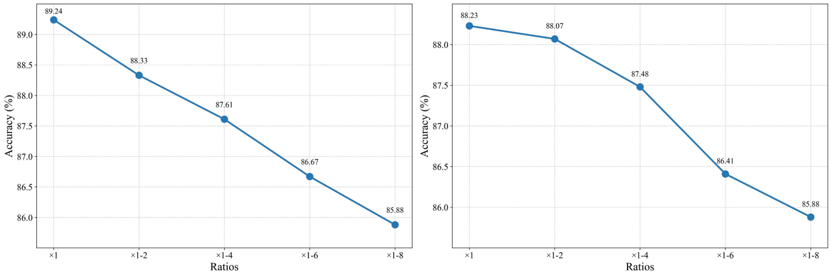
5.4 Ablation Study

Table 3. The result comparison with different training resolution. ×1 means training with original resolution data, $\times 1 - 6$ means training with multi resolution data (×1, ×2, ×4, and ×6).

Train Resolution	×1	×2	×4	×6	×8
×1	**89.24**%	86.17%	77.11%	65.41%	56.22%
×2	86.96%	**88.23**%	79.27%	66.30%	54.50%
×4	81.42%	83.51%	85.20%	73.89%	59.84%
×6	66.72%	70.76%	78.39%	80.35%	69.82%
×8	54.27%	54.60%	64.34%	73.50%	76.86%
$\times 1 - 2$	88.33%	88.07%	81.55%	68.12%	56.19%
$\times 1 - 4$	87.61%	87.48%	**85.30**%	76.37%	64.24%
$\times 1 - 6$	86.67%	86.41%	84.94%	**81.91**%	73.01%
$\times 1 - 8$	85.88%	85.88%	84.25%	80.96%	**77.21**%
DRGFER	**89.24**%	**88.23**%	**85.30**%	**81.91**%	**77.35**%

Table 3 investigates the impact of data augmentation at different resolutions on models trained with various low-resolution data. The results show that when a model is trained using a single resolution, it achieves the best performance only at that specific resolution, which is consistent with the training setting. However, when multiple resolutions are combined for joint training, the accuracy at some lower resolutions can be improved, albeit at the cost of decreased performance at higher resolutions. For example, in Table 3, when we use $\times 1 - 6$ to train the model, it brings a 1.56% improvement compared with only training with single ×6 resolution on the result of testing with ×6. However, the accuracy of ×1, ×2, and ×4 decreases by 2.57%, 1.82%, and 0.71%, respectively, compared to their single-resolution training counterparts.

In this case, similar observations can be made for the ×4 and ×8 downsample ratio, where multi scale low-resolution training leads to accuracy improvements of 0.1% and 0.35%, respectively. These results reinforce the idea that combining multiple resolutions during training can enhance the performance of the model in handling lower-resolution input images. By training jointly on a variety of resolutions, the model becomes more adaptable and robust. However, this increased adaptability comes at the cost of decreased performance at higher resolutions, as the model learns to generalize across different resolutions rather than specializing in high-resolution details.



(a) Recognition accuracy for x1. **(b)** Recognition accuracy for x2.

Fig. 6. Comparison of recognition accuracy.

Furthermore, the more joint resolutions included in the training, the lower the accuracy becomes, as shown in Fig. 6 and Fig. 6b. This phenomenon can be attributed to the fact that models trained on low-resolution images may not be as effective in capturing high-resolution features, leading to poorer performance on high-definition input images. Therefore, while joint training can enhance the model's adaptability to lower-resolution images, it may also compromise its performance at higher resolutions.

The DRGFER proposed in this paper can directly avoid the FER model from facing data of inappropriate resolution, thereby maximizing the recognition effect.

6 Conclusion

In this paper, we proposed a novel method called Dynamic Resolution Guidance for Facial Expression Recognition (DRGFER) to effectively recognize facial expressions in different low-resolution images without compromising the accuracy of the FER model. Our framework consists of two main components: the Resolution Recognition Network (RRN) and the Multi-Resolution Adaptation Facial Expression Recognition Network (MRAFER). Our proposed DRGFER framework demonstrates a practical and effective approach to handle and process facial expression images with varying resolutions.

Limitations and Future Works. While our method can automatically adapt to a variety of low-resolution facial expression recognition tasks, there are still some limitations to be addressed. One limitation is that our definition of low resolution is discrete and not sufficiently detailed, which may lead to inconsistencies between the settings during training and real-world applications. To mitigate this issue, we plan to continue exploring ways to refine the resolution strength within our framework, enabling a more fine-grained and continuous representation of resolution variations.

Another limitation of our current approach is the requirement of multiple independent FER_Blocks, which, although not introducing additional computational overhead, can increase memory usage. To address this, we aim to investigate the establishment of more weight-sharing mechanisms to reduce memory overhead. Additionally, we believe that incorporating the concept of knowledge sharing among the FER_Blocks can further improve the overall performance of our method.

By addressing these limitations and exploring these potential improvements, we strive to develop a more robust, efficient, and adaptable solution for real-world applications.

References

1. Barsoum, E., Zhang, C., Ferrer, C.C., Zhang, Z.: Training deep networks for facial expression recognition with crowd-sourced label distribution. In: Proceedings of the 18th ACM International Conference on Multimodal Interaction, pp. 279–283 (2016)
2. Canal, F.Z., et al.: A survey on facial emotion recognition techniques: a state-of-the-art literature review. Inf. Sci. **582**, 593–617 (2022)
3. Cheng, B., et al.: Robust emotion recognition from low quality and low bit rate video: a deep learning approach. In: 2017 Seventh International Conference on Affective Computing and Intelligent Interaction (ACII), pp. 65–70. IEEE (2017)
4. 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)
5. 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)
6. Hilles, M.M., Naser, S.S.A.: Knowledge-based intelligent tutoring system for teaching mongo database.(2017) (2017)
7. Hu, X., Mu, H., Zhang, X., Wang, Z., Tan, T., Sun, J.: Meta-SR: a magnification-arbitrary network for super-resolution. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 1575–1584 (2019)
8. Jing, W., Tian, F., Zhang, J., Chao, K.M., Hong, Z., Liu, X.: Feature super-resolution based facial expression recognition for multi-scale low-resolution faces. arXiv preprint arXiv:2004.02234 (2020)
9. Lai, W.S., Huang, J.B., Ahuja, N., Yang, M.H.: Deep Laplacian pyramid networks for fast and accurate super-resolution. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 624–632 (2017)

10. Li, S., Deng, W., Du, J.: Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2852–2861 (2017)
11. 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)
12. Liu, Z., Li, L., Wu, Y., Zhang, C.: Facial expression restoration based on improved graph convolutional networks. In: Ro, Y.M., et al. (eds.) MMM 2020. LNCS, vol. 11962, pp. 527–539. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-37734-2_43
13. Lukas, S., Mitra, A.R., Desanti, R.I., Krisnadi, D.: Student attendance system in classroom using face recognition technique. In: 2016 International Conference on Information and Communication Technology Convergence (ICTC), pp. 1032–1035. IEEE (2016)
14. Nan, F., et al.: Feature super-resolution based facial expression recognition for multi-scale low-resolution images. Knowl.-Based Syst. **236**, 107678 (2022)
15. Ou, J., Wu, H.: Efficient human pose estimation with Depthwise separable convolution and person centroid guided joint grouping. In: Peng, Y., et al. (eds.) PRCV 2020. LNCS, vol. 12306, pp. 626–638. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-60639-8_52
16. Paszke, A., et al.: PyTorch: an imperative style, high-performance deep learning library. In: Advances in Neural Information Processing Systems, vol. 32 (2019)
17. Shao, J., Cheng, Q.: E-FCNN for tiny facial expression recognition. Appl. Intell. **51**, 549–559 (2021)
18. Szegedy, C., et al.: Going deeper with convolutions. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1–9 (2015)
19. Tang, J., Zhou, X., Zheng, J.: Design of intelligent classroom facial recognition based on deep learning. In: Journal of Physics: Conference Series. vol. 1168, p. 022043. IOP Publishing (2019)
20. Wu, G., Jiang, J., Liu, X., Ma, J.: A practical contrastive learning framework for single image super-resolution. arXiv preprint arXiv:2111.13924 (2021)
21. Zhang, Y., Liu, T., Long, M., Jordan, M.: Bridging theory and algorithm for domain adaptation. In: International Conference on Machine Learning, pp. 7404–7413. PMLR (2019)
22. 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)
23. Zhang, Y., Tian, Y., Kong, Y., Zhong, B., Fu, Y.: Residual dense network for image super-resolution. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2472–2481 (2018)
24. Zhu, M., et al.: Dynamic resolution network. Adv. Neural. Inf. Process. Syst. **34**, 27319–27330 (2021)

Author Index

A. Antonacopoulos et al. (Eds.): ICPR 2024, LNCS 15332, pp. 423–424, 2025.
https://doi.org/10.1007/978-3-031-78125-4