LNCS 15329

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

Pattern Recognition

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

29 Part XXIX

ICPR 2024 INDIA

IAPR

Springer

MOREMEDIA

Lecture Notes in Computer Science 15329

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 XXIX

Springer

Editors

Apostolos Antonacopoulos
University of Salford
Salford, UK

Subhasis Chaudhuri
Indian Institute of Technology Bombay
Mumbai, India

Rama Chellappa
Johns Hopkins University
Baltimore, MD, USA

Cheng-Lin Liu
Chinese Academy of Sciences
Beijing, China

Saumik Bhattacharya
IIT Kharagpur
Kharagpur, India

Umapada Pal
Indian Statistical Institute Kolkata
Kolkata, India

ISSN 0302-9743 ISSN 1611-3349 (electronic)
Lecture Notes in Computer Science
ISBN 978-3-031-78109-4 ISBN 978-3-031-78110-0 (eBook)
https://doi.org/10.1007/978-3-031-78110-0

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

President's Address

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

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

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

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

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

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

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

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

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

September 2024

Arjan Kuijper
President of the IAPR

Preface

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

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

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

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

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

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

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

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

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

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

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

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

October 2024

Umapada Pal
Josef Kittler
Anil Jain

Organization

General Chairs

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

Program Chairs

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

Publication Chairs

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

Competition Chairs

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

Workshop Chairs

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

Tutorial Chairs

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

Doctoral Consortium Chairs

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

Organizing Chairs

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

Organizing Committee

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

Finance Chairs

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

Awards Committee Chair

Arpan Pal	Tata Consultancy Services, India

Sponsorship Chairs

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

Exhibition and Demonstration Chairs

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

International Liaison, Visa Chair

Balasubramanian Raman	Indian Institute of Technology, Roorkee, India

Publicity Chairs

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

Women in ICPR Chairs

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

Event Manager

Alpcord Network

Track Chairs – Artificial Intelligence, Machine Learning for Pattern Analysis

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

Track Chairs – Computer and Robot Vision

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

Track Chairs – Image, Speech, Signal and Video Processing

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

Track Chairs – Biometrics and Human Computer Interaction

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

Track Chairs – Document Analysis and Recognition

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

Track Chairs – Biomedical Imaging and Bioinformatics

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

Metareviewers (Conference Papers and Competition Papers)

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

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

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

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

Reviewers (Competition Papers)

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

Reviewers (Conference Papers)

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

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

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

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

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

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

Jinwoo Ahn
Jinwoo Choi
Jinyang Liu
Jinyu Tian
Jionghao Lin
Jiuding Duan
Jiwei Shen
Jiyan Pan
Jiyoun Kim
João Papa
Johan Debayle
John Atanbori
John Wilson
John Zhang
Jónathan Heras
Joohi Chauhan
Jorge Calvo-Zaragoza
Jorge Figueroa
Jorma Laaksonen
José Joaquim De Moura Ramos
Jose Vicent
Joseph Damilola Akinyemi
Josiane Zerubia
Juan Wen
Judit Szücs
Juepeng Zheng
Juha Roning
Jumana H. Alsubhi
Jun Cheng
Jun Ni
Jun Wan
Junghyun Cho
Junjie Liang
Junjie Ye
Junlin Hu
Juntong Ni
Junxin Lu
Junxuan Li
Junyaup Kim
Junyeong Kim
Jürgen Seiler
Jushang Qiu
Juyang Weng
Jyostna Devi Bodapati
Jyoti Singh Kirar

Kai Jiang
Kaiqiang Song
Kalidas Yeturu
Kalle Åström
Kamalakar Vijay Thakare
Kang Gu
Kang Ma
Kanji Tanaka
Karthik Seemakurthy
Kaushik Roy
Kavisha Jayathunge
Kazuki Uehara
Ke Shi
Keigo Kimura
Keiji Yanai
Kelton A. P. Costa
Kenneth Camilleri
Kenny Davila
Ketan Atul Bapat
Ketan Kotwal
Kevin Desai
Keyu Long
Khadiga Mohamed Ali
Khakon Das
Khan Muhammad
Kilho Son
Kim-Ngan Nguyen
Kishan Kc
Kishor P. Upla
Klaas Dijkstra
Komal Bharti
Konstantinos Triaridis
Kostas Ioannidis
Koyel Ghosh
Kripabandhu Ghosh
Krishnendu Ghosh
Kshitij S. Jadhav
Kuan Yan
Kun Ding
Kun Xia
Kun Zeng
Kunal Banerjee
Kunal Biswas
Kunchi Li
Kurban Ubul

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

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

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

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

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

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

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

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

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

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

Contents – Part XXIX

Optimizing Personalized Robot Actions with Ranking of Trajectories 1
Hao Huang, Yiyun Liu, Shuaihang Yuan, Congcong Wen, Yu Hao, and Yi Fang

Synthesizing Operationally Safe Controllers for Human-in-the-Loop Human-in-the-Plant Hybrid Close Loop Systems 17
Ayan Banerjee, Imane Lamrani, and Sandeep K. S. Gupta

Multi-frequency Fine-Grained Matching for Audio-Visual Segmentation 36
Yinhao Zhang, Tianyang Xu, Xiao-Jun Wu, Shaochuan Zhao, and Josef Kittler

Confidence-Guided Feature Alignment for Cloth-Changing Person Re-identification 51
Sirong Huang and Huicheng Zheng

Fair Latent Representation Learning with Adaptive Reweighing 65
Puspita Majumdar, Raghav Sharma, Rohit Bhattacharya, and Balraj Prajesh

FineFACE: Fair Facial Attribute Classification Leveraging Fine-Grained Features 81
Ayesha Manzoor and Ajita Rattani

One-Factor Cancelable Biometric Template Protection Scheme for Real-Valued Features 98
Ruoqi Zhang, Peisong Shen, Kewei Lv, and Chi Chen

Multi-teacher Invariance Distillation for Domain-Generalized Action Recognition 116
Jongmin Shin, Abhishek Maiti, Yuliang Zou, and Jinwoo Choi

ALS-HAR: Harnessing Wearable Ambient Light Sensors to Enhance IMU-Based Human Activity Recogntion 133
Lala Shakti Swarup Ray, Daniel Geißler, Mengxi Liu, Bo Zhou, Sungho Suh, and Paul Lukowicz

Spatio-Temporal Domain-Aware Network for Skeleton-Based Action Representation Learning 148
Jiannan Hu, Cong Wu, Tianyang Xu, Xiao-Jun Wu, and Josef Kittler

Project and Pool: An Action Localization Network for Localizing Actions in Untrimmed Videos 164
Himanshu Singh, Avijit Dey, Badri Narayan Subudhi, and Vinit Jakhetiya

Multi-teacher Importance Preserving Knowledge Distillation for Early Violence Prediction 179
Suvramalya Basak, Aditya Vaishy, and Anjali Gautam

Improving Temporal Action Segmentation and Detection with Hierarchical Task Grammar 196
Qiu Yihui and Deepu Rajan

Hybrid Human Action Anomaly Detection Based on Lightweight GNNs and Machine Learning 212
Miao Feng and Jean Meunier

Zero-Shot Spatio-Temporal Action Detection by Enhancing Context-Relation Capability of Vision-Language Models 229
Yasunori Babazaki, Takashi Shibata, and Toru Takahashi

Nonlinear Progressive Denoising: A Universal Regularized Denoising Strategy for Low PSNR Images 245
Ziyu Wu and Silong Peng

Unified-EGformer: Exposure Guided Lightweight Transformer for Mixed-Exposure Image Enhancement 260
Eashan Adhikarla, Kai Zhang, Rosaura G. VidalMata, Manjushree Aithal, Nikhil Ambha Madhusudhana, John Nicholson, Lichao Sun, and Brian D. Davison

Composite Concept Extraction Through Backdooring 276
Banibrata Ghosh, Haripriya Harikumar, Khoa D. Doan, Svetha Venkatesh, and Santu Rana

Guided SAM: Label-Efficient Part Segmentation 291
Sabina B. van Rooij and Gertjan J. Burghouts

Multidimensional Cross-Reconstructed Networks for Few-Shot Fine-Grained Image Classification 307
Yu Cheng, Bo Li, Penghao Jia, Aoxiang Ning, and Jinhong He

Multiplicative RMSprop Using Gradient Normalization for Learning Acceleration 322
Manos Kirtas, Nikolaos Passalis, and Anastasios Tefas

TADIL: Task-Agnostic Domain-Incremental Learning Through Task-ID Inference Using Transformer Nearest-Centroid Embeddings 337
Gusseppe Bravo-Rocca, Peini Liu, Jordi Guitart, Ajay Dholakia, and David Ellison

Evidential Federated Learning for Skin Lesion Image Classification 354
Rutger Hendrix, Federica Proietto Salanitri, Concetto Spampinato, Simone Palazzo, and Ulas Bagci

Adaptive Text Feature Updating for Visual-Language Tracking 366
Xuexin Liu, Zhuojun Zou, and Jie Hao

TrajDiffuse: A Conditional Diffusion Model for Environment-Aware Trajectory Prediction 382
Qingze Tony Liu, Danrui Li, Samuel S. Sohn, Sejong Yoon, Mubbasir Kapadia, and Vladimir Pavlovic

Principal Graph Neighborhood Aggregation for Underwater Moving Object Detection 398
Meghna Kapoor, Badri Narayan Subudhi, Vinit Jakhetiya, and Ankur Bansal

Dynamic Loss Decay Based Robust Oriented Object Detection on Remote Sensing Images with Noisy Labels 413
Guozhang Liu, Ting Liu, Mengke Yuan, Tao Pang, Guangxing Yang, Hao Fu, Tao Wang, and Tongkui Liao

Environment-Independent Fusion for Robust Object Detection in Adverse Environments 429
Wenlong Zhong, Yunfei Zhang, and Si Wu

Transformer-Based RGB and LiDAR Fusion for Enhanced Object Detection 445
Reza Sadeghian, Niloofar Hooshyaripour, and WonSook Lee

Author Index 461

Optimizing Personalized Robot Actions with Ranking of Trajectories

Hao Huang[1](✉), Yiyun Liu[2], Shuaihang Yuan[1], Congcong Wen[1], Yu Hao[1], and Yi Fang[1]

[1] Embodied AI and Robotics (AIR) Lab, NYU Abu Dhabi, New York University Abu Dhabi, Abu Dhabi, UAE
hh1811@nyu.edu

[2] Tandon School of Engineering, New York University, New York, USA

Abstract. Intelligent robots designed for real-world human interactions need to adapt to the diverse preferences of individuals. Preference-based Reinforcement Learning (PbRL) offers promising potential to teach robots personalized behaviors by learning through interactions with humans, eliminating the need for intricate, manually crafted reward functions. However, the current PbRL approaches are hampered by sub-optimal feedback efficiency and limited exploration within state and reward spaces, resulting in subpar performance in complex interactive tasks. To enhance the effectiveness of PbRL, we integrate prior task knowledge into the PbRL framework. Subsequently, we develop a reward model based on ranking a set of multiple robot trajectories. This acquired reward is then utilized to refine the robot's policy, ensuring alignment with human preferences. To validate our method, we showcase its versatility in different human-robot assistive tasks. The experimental results demonstrate that our approach offers a useful, effective, and broadly applicable solution for personalized human-robot interaction.

Keywords: Preference-based reinforcement learning (PbRL) · Human-robot interaction · Multiple trajectory ranking · Assistive Gym

1 Introduction

Recent advances in artificial intelligence and robotics have laid the foundation for the emergence of interactions between humans and robots [9,17]. As stated in [25], the primary objective of robots engaged in human-centered interaction is to improve human security, comfort, and autonomy. In order to accomplish this objective, robots are constructed with the purpose of engaging in ongoing cooperation with individuals within uncontrived settings, wherein they are required to deal with the intricacies posed by a diverse array of human actions and behaviors. To effectively cater to the multifaceted and personalized requirements of human users, it is imperative for robots to possess the capability to adjust and conform to complicated and unique patterns of human behaviors.

H. Huang and Y. Liu—Indicates equal contribution.

A. Antonacopoulos et al. (Eds.): ICPR 2024, LNCS 15329, pp. 1–16, 2025.
https://doi.org/10.1007/978-3-031-78110-0_1

Reinforcement Learning (RL), an example of interactive machine learning, presents a viable solution to tailor robot behaviors for human-robot interactions [8,22]. By leveraging the inherent capabilities of deep neural networks as universal function approximators [16], Deep Reinforcement Learning (DRL) has demonstrated remarkable achievements in various domains characterized by complex problem-solving scenarios. Notable examples include but are not limited to, the triumph of AlphaGo in the realm of board games [38] and the successful navigation of Atari electronic games [28]. Moreover, DRL has also made significant contributions to the field of robotics. However, the effectiveness of these approaches is intricately dependent on the meticulous design of a reward function. Regrettably, a plethora of tasks demonstrate complex and elaborate objectives, leading to a disparity between the reward function manually designed by humans and the true underlying reward function. In the context of human-robot interactions, the incorporation of human or user preference is still insufficient [29]. This deficiency leads to the inability to effectively guide robots toward behaviors that are considered desirable, safe, and aligned with our established human values.

Preference-based Reinforcement Learning (PbRL) is being considered as a promising alternative to address the need for manually designed rewards [7,12,46]. This approach involves learning from non-numeric feedback in sequential domains [47]. In contrast to the conventional approaches of optimizing for a pre-determined long-term reward, the robots in PbRL utilize qualitative feedback. This feedback, often reflecting human preferences, provides information about *two* distinct robot trajectories or demonstrations [48,49]. The purpose of this feedback is to guide the robots' actions toward alignment with human inclinations. A prominent technique in the field of PbRL involves the initial acquisition of a reward function through human feedback. Subsequently, this is followed by an optimization process with respect to the acquired reward function, as described by [7]. However, a notable limitation of the existing PbRL algorithms lies in their lack of efficiency, primarily due to their attempt to acquire a reward function only through *binary* human feedback. The task of achieving comprehensive coverage of the state space becomes increasingly challenging when exploration is exclusively guided by human preferences. As a result, the dependence on binary human feedback presents difficulties in training robots to perform complex interactive tasks effectively while also aligning with human preferences.

Inspired by [39] in which the existing *pairwise* comparison methodology proposed by Bradley-Terry [5] is extended to encompass the inclusion of preference rankings for *multiple* trajectories. It is assumed that a collection of ranked trajectories is available. The aim is to learn a reward function by analyzing these trajectories in a way that aligns well with human preference. This is in direct opposition to traditional PbRL approaches that prioritize solely the optimal trajectory or the conventional Reinforcement Learning from Human Feedback (RLHF) approach that relies exclusively on pairwise comparisons to instruct the reward model learning process [7,18,40]. Our approach is positioned to effectively replicate the objectives of human preference alignment with greater precision. Moreover, the implementation of our model requires the incorporation of a differentiable contrastive loss function [39] to replace the cross-entropy loss function utilized in previous works using pairwise trajectories [7,18,24]. To validate the effectiveness of our method, we conducted experiments in a physical simulation environment, *i.e.*, Assistive Gym [11], focusing on four distinct tasks, *i.e.*, scratching, bathing, feeding, and drinking, that involve human-robot interaction. The experimental results demonstrate a significant improvement in the effectiveness of PbRL approaches, specifically on-policy PrefPPO [21], when employing our proposed method.

The rest of the paper is organized as follows. Following a discussion of the existing literature in Sect. 2, we proceed to detail the proposed approach in Sect. 3. In Sect. 4, experiments are conducted with human-robot interaction and the results substantiate the efficacy of the proposed approach. The conclusions are presented in Sect. 5.

2 Related Work

Human-Robot Interaction and Collaboration. The field of human-robot interaction centers its attention on the dynamics of interaction between human beings and robots [31]. To facilitate the robot's adherence to human preference, it is imperative that the robot possesses the capability to detect, interpret, and react to human states, actions, intentions, and even emotions. Research related to human-robot interaction includes the enhancement of robot social acuity through the utilization of visual perception of individuals [43] and the integration of natural language processing to facilitate more intuitive and effective communication between humans and robots [50]. In the field of human-robot collaboration, robotic systems are usually designed to function in a manner that complements or enhances the desired objective of human individuals [3]. Collaboration with a robotic entity has the potential to improve task efficiency and increase work productivity, thus reducing the occurrence of errors. The implementation of personalized collaborative plans is carried out for the purpose of robot-assisted dressing [19]. Furthermore, it can contribute to the improvement of human safety by mitigating the adverse effects of repetitive strain and minimizing the likelihood of injury [14,45]. In recent studies, RL has demonstrated its capacity to facilitate personalized human-robot interaction by optimizing the parameters of an interaction model [10,32,42,51].

Preference-Based Reinforcement Learning. PbRL is a sub-realm of reinforcement learning in which the robot learns optimal policies based on human or expert preference feedback rather than explicit scalar reward functions. Preference-based reinforcement learning has been extended to address complex challenges, such as Atari games [18] and robotic tasks [49], by incorporating deep learning models. One widely used approach involves learning a reward function from human feedback and subsequently optimizing a learnable model to maximize the learned reward [1]. The main obstacles in PbRL pertain to the optimization of sample and feedback efficiency, as the acquisition of human feedback in real-world scenarios is sometimes expensive. Previous studies have been conducted to investigate the effectiveness of using paired preference input from humans for training robots [12,46]. In recent advances, an off-policy preference-based RL algorithm has been introduced that enhances both these efficiencies by relabeling historical experiences and employing unsupervised pretraining [21]. Furthermore, the synthesis of expert demonstrations with pairwise preferences has been validated as a powerful mechanism to improve the efficiency of PbRL [18,33]. The integration of uncertainty in reward functions has led to the development of effective exploration approaches [23]. Alternatively, in an effort to reduce the reliance on human feedback for query optimization without performance degradation, previous research has explored the use of preference predictors that rely on pseudo-preference labels [34,49]. Distinct from the previous works in which pairwise demonstrations are used for preference query, we adapt [30,39] to rank multiple demonstrations to more effectively explore an action space while maintaining alignment with human preference.

Luce's Choice Axiom for Human Preference. Luce's choice axiom [26], a foundational principle in decision theory, posits that the probability of choosing an item from a set of alternatives is proportional to the item's utility (*i.e.*, usually a numerical value) relative to the sum of the utilities of all available options. This axiom has been instrumental in PbRL [13,35], where it is used to model the selection of probabilistic actions based on estimated utilities to reflect human preferences. The Bradley-Terry model [5] is a special case of Luce's choice axiom for pairwise comparisons, determining the probability of preferring one item over another based on their relative strengths. Biyik and Sadigh [4] shows that diversity is important when optimizing a batch of pairwise preference queries. The Plackett-Luce ranking model [2,27] generalizes Luce's choice axiom to handle complete and partial rankings of multiple items, modeling preferences in situations where items are ranked rather than simply chosen or compared in pairs. In the following research, for instance, T-REX [6] uses ranked trajectories to learn reward functions, and Myers *et al.* [30] uses ranked demonstrations to learn multi-modal reward functions.

3 Methods

In Fig. 1, we present a pipeline to learn a reward function based on human preferences using a ranking of multiple trajectories within a reinforcement learning framework. Initially, a set of reward models $\{r_\phi\}$ is trained using the ranking of preferences queried from historical experiences. Then, the learned reward function is combined with a sketchy reward r_{task} [24] to form a total reward r_{total}. This total reward is finally used to train the policy $\pi_\theta(a|s)$, thereby optimizing the robot's actions within the environment.

3.1 Reinforcement Learning with Pairwise Trajectories

Reinforcement learning is a paradigm in which a robot learns through interactions with its environment [41]. At each time step t, the robot observes a state s_t and selects an action a_t according to its policy $\pi(a_t|s_t)$. In standard RL, the environment assigns a numerical reward $r(s_t, a_t)$, with the robot's objective being to maximize the cumulative discounted return $G_t = \sum_{k=0}^{T} \gamma^k r(s_{t+k}, a_{t+k})$ with the horizontal length of T.

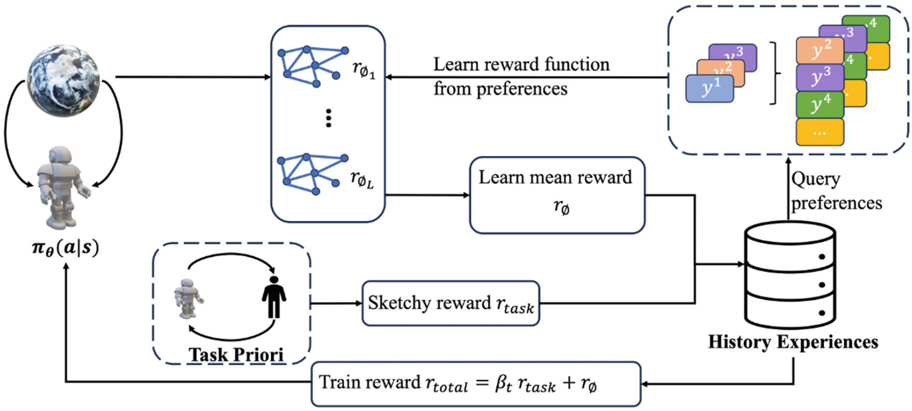


Fig. 1. An illustration of our method (adapted from [24] and [39]). We follow the same learning scheme of [24], but instead the robot is guided by the sum of a learned mean reward $\hat{r}_\phi$, which is optimized by the ranking of *multiple* trajectories to align with human preferences.

Crucially, in this study, we do not assume the direct availability of such a numerical reward. Instead, we postulate that the presence of a human possesses particular intentions or preferences for the task of the robot and conveys these preferences via a comparison of *pairwise* trajectory segments of the robot's actions, favoring those that align more closely with the desired objective. Formally, a single behavior trajectory is denoted as $\tau = \{(s_1, a_1), (s_2, a_2), \cdots, (s_T, a_T)\}$, and the robot demonstrates two trajectories (τ^i, τ^j) to query humans for preference. Human tells which trajectory is preferable through binary feedback, *i.e.*, $z = (\tau^i \succ \tau^j)$ where τ^i is preferred, or $z = (\tau^j \succ \tau^i)$ vise versa, or $z = (\tau^i = \tau^j)$ that these two trajectories are preferred equally. To learn the reward function $\hat{r}_\phi(s_i, a_i)$ which is parametrized by a deep neural network with parameters ϕ to match the preference z received from humans, in previous work [24], the Bradley-Terry [5] model is adopted. Specifically, given a pair of trajectories along with the corresponding human preference $(\tau^i, \tau^j, z = (\tau^i \succ \tau^j))$, we first calculatethe probability that the reward for τ^i

is higher than τ^j [18,33], based on the intuition that the preferred trajectory is expected to yield a higher cumulative reward[1] [24]:

$$P_\phi[\tau^i \succ \tau^j] = \frac{\exp \sum_{t=0}^{T} \hat{r}_\phi(s_t^i, a_t^i)}{\exp \sum_{t=0}^{T} \hat{r}_\phi(s_t^i, a_t^i) + \exp \sum_{t=0}^{T} \hat{r}_\phi(s_t^j, a_t^j)} \ , \tag{1}$$

and then a supervised classification learning scheme is applied to train the reward function to match with human preference [18,24]:

$$\begin{aligned} \mathcal{L}_\phi = -\mathbb{E}_{(\tau^i,\tau^j,z)\sim\mathcal{D}} \Big[& \mathbb{I}[z = (\tau^i \succ \tau^j)] \log P_\phi[\tau^i \succ \tau^j] \\ & + \mathbb{I}[z = (\tau^j \succ \tau^i)] \log P_\phi[\tau^j \succ \tau^i] \Big] \ , \end{aligned} \tag{2}$$

where $\mathbb{I}[\cdot]$ represents an indicator function and $\mathcal{D}$ is dataset storing trajectories with their preference feedbacks. Once the reward function $\hat{r}_\phi$ has been optimized according to human preferences, the policy $\pi_\theta(s_i, a_i)$ which is also parametrized by a deep neural network with parameters θ can be learned using any conventional RL approach with the reward function $\hat{r}_\phi$.

3.2 Reinforcement Learning with Multiple Trajectories

Instead of employing the Bradley-Terry paired comparison [5] to fit a reward model to human preferences, we consider encompassing comparisons within preference rankings of multiple numbers of robot behavior trajectory segments [39]. Instead of demonstrating a pair of behavior trajectories (y^i, y^j) to query humans for preference, the robot demonstrates M $(M > 2)$ behavior trajectories for a query. The human then annotates the indices of the order of M trajectories as $\boldsymbol{Y} = (y^1, y^2, \cdots, y^M)$, where y^1 represents the index of the trajectory with the *highest* preference score, *i.e.*, y^1 is the most preferred trajectory. Furthermore, all trajectories in $\boldsymbol{Y}$ satisfy $y^i \succ y^j, \forall i < j$. We store these preference feedbacks in a dataset $\mathcal{D}$ as $\boldsymbol{Y} \sim \mathcal{D}$. To perform PbRL based on the preference for ranked trajectories as described above, we follow the subsequent steps.

Query Selection. PbRL aims to train robots to exhibit human-desired behaviors using minimal preference feedback, initiating by selecting trajectories for human queries. Throughout the training, historical trajectories are stored in a buffer $\mathcal{B}$, and the robot generates N_{query} trajectories per feedback session to gather human preferences. The choice of query strategy is critical to minimize human effort. Uniform sampling, the process of randomly selecting N_{query} groups of M trajectories from the buffer $\mathcal{B}$, is a straightforward approach. Alternatively, ensemble-based sampling, a more sophisticated method, seeks to optimize information gain by choosing trajectories with the largest variance in predictions across multiple reward models, as discussed in previous works [7,18,21,24]. We evaluated both selection strategies in our experiment in Sect. 4. After querying human preferences, we store the ranked trajectories in $\mathcal{D}$.

[1] The probability $P_\phi[\tau^j \succ \tau^i]$ is calculated in a similar way for human preference $\big(\tau^i, \tau^j, z = (\tau^j \succ \tau^i)\big)$.

Rewards Learning from Preference Ranking. As discussed above, $\boldsymbol{Y} = (y^1, y^2, \cdots, y^M)$ indicates $y^1 \succ y^2 \succ \cdots \succ y^M$. We further define the partial order between y^1 and all the other candidates behind it as $\boldsymbol{\tau}[y^1, y^2 : y^M] = y^1 \succ \{y^2, \cdots, y^M\}$, then the objective of Bradley-Terry comparison for multiple ranked trajectories becomes [30]:

$$P(\boldsymbol{\tau}[y^1, y^2 : y^M]) = \frac{\exp \sum_{t=0}^{T} \hat{r}_\phi(s_t^1, a_t^1)}{\sum_{i=1}^{M} \exp \sum_{t=0}^{T} \hat{r}_\phi(s_t^i, a_t^i)} \ . \tag{3}$$

Furthermore, it is essential to notice that the objective in Eq. 3 does not fully leverage the rankings $y^1 \succ y^2 \succ \cdots \succ y^M$, as it only characterizes $y^1 \succ \{y^2, \cdots, y^M\}$, disregarding the remaining $M-2$ valuable rankings, such as $y^2 \succ \{y^3, \cdots, y^M\}$ and $y^{M-1} \succ y^M$. As each relative ranking of any pair of trajectories in $\boldsymbol{Y}$ reflects human preference from different aspects, to capture human preference as completely as possible and speed up the convergence of the reward function learning process, we adopt an extension to Eq. 3 as [30,39]:

$$P(\boldsymbol{\tau}[y^1, y^2, \cdots, y^M]) = \prod_{k=1}^{M-1} P(\boldsymbol{\tau}[y^k, y^{k+1} : y^M]) = \prod_{k=1}^{M-1} \frac{\exp \sum_{t=0}^{T} \hat{r}_\phi(s_t^k, a_t^k)}{\sum_{i=k}^{M} \exp \sum_{t=0}^{T} \hat{r}_\phi(s_t^i, a_t^i)} \ . \tag{4}$$

If $M \to \infty$, then Eq. 4 is able to exhaustively explore all possible trajectories and annotate y^1 as the most desired one, and thus perfectly alignments with human preference. On the contrary, if $M = 2$, Eq. 4 degenerates into Eq. 1 for pairwise Bradley-Terry comparison. Thereafter, the reward function $\hat{r}_\phi$ is optimized to minimize the loss $\mathcal{L}_\phi$ [39]:

$$\mathcal{L}_\phi = -\mathbb{E}_{\boldsymbol{Y} \sim \mathcal{D}} \Big[\sum_{k=1}^{M-1} \log \frac{\exp \sum_{t=0}^{T} \hat{r}_\phi(s_t^k, a_t^k)}{\sum_{i=k}^{M} \exp \sum_{t=0}^{T} \hat{r}_\phi(s_t^i, a_t^i)} \Big] \ . \tag{5}$$

Also, notice that Eq. 5 is an extension of Eq. 2 for multiple trajectories and shares a similar form as contrastive loss in [15,20] which is implementation-friendly.

Discussion. Using ranked trajectories based on human preference to guide the learning process of reward functions has been explored in previous literature, such as T-REX [6] and Myers *et al.* [30]. However, T-REX repeatedly applies the binary loss defined in Eq. 1 and Eq. 2 to each pair in a group of ranked trajectories. Instead, we adopt a simplified ranking model as used in [30] in which we assign the mixing coefficients of each reward function parameterized by a neural network to be the same value. Such a ranking model has also been used in [39] as an effective supervised fine-tuning algorithm to fine-tune large language models for better alignment with human preferences, enhancing the traditional pairwise contrast method to handle preference rankings of inputs with varying lengths.

Policy Optimization. Once optimized based on human preferences using Eq. 5, the reward function $\hat{r}_\phi$ transforms PbRL into a standard RL challenge,

Algorithm 1. PbRL Using Ranking of Multiple Trajectories (adapted from [24])

```
Initialize a set of reward functions $\{\hat{r}_{\phi_l}\}_{l=1}^{L}$ and robot action policy $\pi_\theta$
Define a task-specific sketchy reward $\hat{r}_{task}$
Initialize human feedback frequency $P$
Initialize buffer $\mathcal{B}$ to store robot trajectories
Initialize dataset $\mathcal{D}$ to store preference feedback
for each training time step $t$ do
    // INTERACTION WITH THE ENVIRONMENT
    Collect $s_{t+1}$ by running $a_t \sim \pi_\theta(a_t|s_t)$
    Store transition $\{s_t, a_t, s_{t+1}, \hat{r}_{task}(s_t, a_t)\}$ into buffer $\mathcal{B}$
    if $t\%P == 0$ then
        // QUERY SELECTION
        Select $N_{query}$ groups of $M$ trajectories from buffer $\mathcal{B}$
        Query human preference indices $\boldsymbol{Y}$ and store them as $\mathcal{D} \leftarrow \mathcal{D} \cup \{(\boldsymbol{\tau}_i, \boldsymbol{Y}_i)\}_{i=1}^{N_{query}}$
        // REWARD LEARNING
        Optimize reward functions $\{\hat{r}_{\phi_l}\}_{l=1}^{L}$ using Eq. 5
    end if
    // POLICY OPTIMIZATION
    for each optimization iteration do
        Sample a mini-batch $\{s_j, a_j, \hat{r}_{task}(s_j, a_j)\}_{j=1}^{N}$ from buffer $\mathcal{B}$
        Compute $\{\hat{r}_\phi(s_j, a_j) = \frac{1}{L}\sum_{l=1}^{L} \hat{r}_{\phi_l}(s_j, a_j)\}_{j=1}^{N}$, update $\beta_t$ using Eq. 7 and
compute $r_{total}$ using Eq. 6
        Train policy $\pi_\theta$ using reward $r_{total}$
    end for
end for
```

allowing the use of any existing RL algorithms to train the policy. For instance, in algorithms such as PrefPPO [21], the conventional reward function is simply replaced with the learned $\hat{r}_\phi$ [7]. However, preference feedback, whether from pairwise or multiple trajectories, yields less information than direct numerical rewards. Therefore, PbRL algorithms are sometimes less efficient than conventional RL algorithms that explicitly utilize task-specific manually designed numerical rewards. In long episodic human-robot interactions, the problem is more severe to assign rewards to the actions at each step. Furthermore, it is also known that PbRL encounters difficulties in achieving comprehensive coverage of state and action spaces by random exploration, particularly in complex robotic tasks involving high-dimensional spaces, as shown in the Assistive Gym physical simulation environment [11]. In addition, as suggested in [44], using preference-based RL in Assistive Gym tasks often suffers from misidentification of rewards. To resolve these issues, we also incorporate task-specific prior knowledge into our PbRL model as proposed in [24]. The primary concept is to separate the task from human preference and additionally establish a *sketchy* reward function to convey the desired behavior exclusively for the task. Once the robot has acquired a comprehensive understanding of the task by a few iterations of trial and error, we employ the reward functionlearned through PbRL as detailed above to opti-

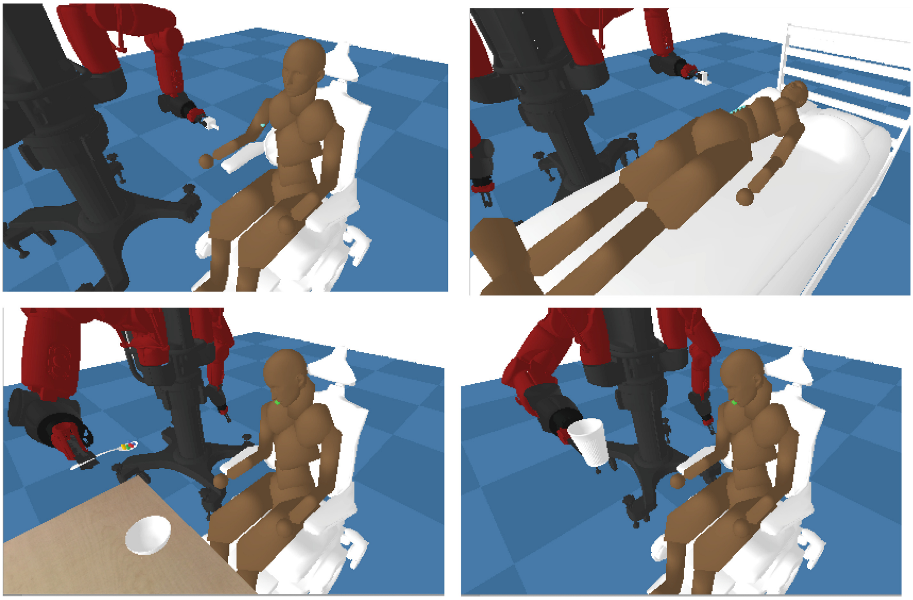


Fig. 2. Visualization of experiments on four tasks from Assistive Gym [11]: itch scratching (top-left), bed bathing (top-right), feeding (bottom-left), and drinking water (bottom-right).

mize the robot's policy. Therefore, the overall reward can be realized as shown in Eq. 6 [24]:

$$r_{total}(s_t, a_t) = \hat{r}_\phi(s_t, a_t) + \beta_t \hat{r}_{task}(s_t, a_t) \ , \tag{6}$$

where the task reward rate, $\beta_t \geq 0$, mediates the trade-off between task completion and alignment with human preferences at training time step t. Given the inaccessibility of the true task reward and the imprecision of the manually designed reward function $\hat{r}_{task}$, we employ a decreasing reward rate over the training process as adopted in [24]:

$$\beta_t = \frac{T_s - t}{T_s}\beta_0 \tag{7}$$

where the hyperparameter T_s is the total training steps and β_0 is the initial value of the task reward rate. The proposed preference-based reinforcement learning using multiple ranked trajectories is outlined in Algorithm 1.

4 Experiments

To assess the effectiveness of our method in personalized human-robot interaction, we conducted experiments within Assistive Gym [11], a physical simulation environment tailored for assistive robotics.

4.1 Setups

Our proposed method is evaluated on four Assistive-Gym tasks, as shown in Fig. 2, including:

Table 1. Activation vector $\boldsymbol{\alpha}$ in different simulation environments.

Bed Bathing	$C_v(s_i)$	$C_f(s_i)$	$C_{hf}(s_i)$	$C_{fd}(s_i)$	$C_{fdv}(s_i)$	$C_d(s_i)$	$C_e(s_i)$
Itch Scratching	1	1	1	0	0	1	1
Bed Bathing	1	1	1	0	0	1	1
Feeding	1	1	1	1	1	1	1
Drinking Water	1	1	1	1	1	1	1

- **Itch Scratching.** A robot equipped with a small scratching tool must reach a specific target spot on the right arm of a person. It earns a reward $r_R(s_i)$ for positioning its end effector near the target location while applying a force of less than $10N$.
- **Bed Bathing.** A robot uses a washcloth to clean the right arm of a person resting in bed. It gains a reward $r_R(s_i)$, for moving nearer to the person's body and effectively wiping along the surface of the right arm.
- **Feeding.** A robot holding a spoon filled with food-like spheres feeds a person by carefully guiding the spoon to the mouth of a person without spilling. It gains a reward $r_R(s_i)$ for moving nearer to the person's mouth and successfully placing the spoon inside.
- **Drinking Water.** A robot assists a person by holding a cup filled with water-like beads, guiding it toward the person's mouth, tilting it, and pouring the liquid into their mouth. The robot earns a reward $r_R(s_i)$ for each step that brings the cup nearer to the mouth and successfully delivers the water.

In these tasks, the policy training reward $r(s_i)$ has two components: the task success reward $r_R(s_i)$ as described above, and the human preference reward $r_H(s_i)$ defined as [24]:

$$r_H(s_i) = -\boldsymbol{\alpha} \cdot \boldsymbol{\omega} \odot [C_v(s_i), C_f(s_i), C_{hf}(s_i), C_{fd}(s_i), C_{fdv}(s_i), C_d(s_i), C_e(s_i)] \ , \tag{8}$$

where $\boldsymbol{\alpha}$ is a binary vector, indicating the active human preferences for a specific task, as presented in Table 1, whereas $\boldsymbol{\omega}$ is a vector of preference weights. In this study, we set $\boldsymbol{\omega} = [0.25, 0.3, 0.1, 2.5, 10.0, 0.5, 0.5]$ and the penalty terms are defined as [24]:

- $C_v(s_i)$: cost for the velocities of the high robot's end effector.
- $C_f(s_i)$: force applied away from the target location.
- $C_{hf}(s_i)$: high forces away from the target ($> 10N$).

- $C_{fd}(s_i)$: spilling food/water out on the person.
- $C_{fdv}(s_i)$: food/water fed into mouth at high velocities.
- $C_d(s_i)$: cost for the distance from the target location to the robot's end effector.
- $C_e(s_i)$: reward for successfully finishing the task.

To evaluate PbRL's efficacy in learning from non-numeric feedback, we assume the robot cannot directly observe the true reward $r(s_i)$. Instead, it learns from preference feedback provided by a scripted human teacher [7,18,21,24]. For better policy optimization, following the approach outlined in [24], we integrate task-specific knowledge $\hat{r}_{task}$ with the learned reward model to improve task exploration. In our simulations, we define $\hat{r}_{task} = -|d|^2$, where d represents the distance between the target location and the end effector of the robot. We use a linear decay type for β_t, as indicated in Eq. 7, starting with an initial value of $\beta_0 = 1$.

4.2 Performance and Analysis

To evaluate the performance of our method, we compared it with the state-of-the-art PbRL methods, *e.g.*, PrefPPO [21] and Decoupled PrefPPO [24], in a feeding task. Our evaluation criteria encompass three key indicators:

- **Episode True Return:** calculated as $r(s_i)$.
- **Task Success Rate:** measures task completion.
- **Episode Preference Return:** calculated as $r_H(s_i)$.

We compared our method with two baselines: RL (*i.e.*, PPO [37]) with the true reward and Decoupled PrefPPO, which updates its reward model using paired trajectories. Our main goal was not to outperform PPO with the true reward but to achieve comparable performance while aligning with human preference. To select trajectories from the buffer for human preference feedback, we explored two sampling methods: uniform sampling and ensemble-based sampling with $L = 3$ ensemble members. We train the reward model using Eq. 5 and train the action policy using Eq. 6 as used in [24].

The results are shown in Fig. 3. In the feeding task, with uniform sampling, our method nearly matches PPO with the true reward, outperforming the pair-query approach. We notice that our methods exhibit slightly lower performance with ensemble sampling. We conjecture that it may be the case that the batch to be ranked is all similarly uncertain without taking diversity into account, leading to low information gain over the entire ranking. Moreover, uniform sampling simplifies the process and reduces runtime compared to ensemble sampling. These findings emphasize the potential of our methods to improve preference-based RL in complex interactive tasks while streamlining trajectory selection.

4.3 Generalization Ability of Our Method

The need for adaptability across different application environments is evident, and excelling in the feeding task alone does not ensure the effectiveness of our

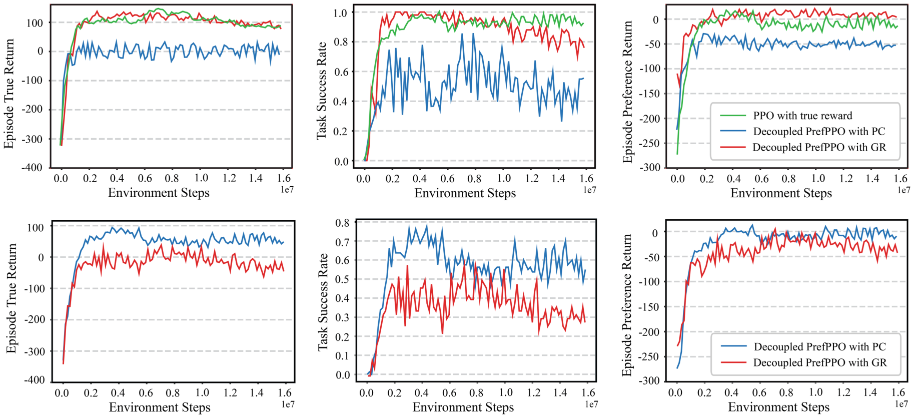


Fig. 3. The learning curves in the feeding task with uniform sampling (top row) and ensembling sampling (bottom row). ‘PPO with true reward’ means PPO receives true reward score directly through the environment; ‘Decoupled PrefPPO with PC’ means using a *pair* of trajectory to query human to update the label of the reward model; ‘Decoupled PrefPPO with GR’ means using a *group* of trajectory ranking to query human to update the label of the reward model.

Table 2. Different group ranking size M in the feeding task.

M	2	5	10	15
Best Task Success Rate	0.88	1.00	0.73	0.69

method. To measure the generalization of our method, we extended it to different assistive gym scenarios, including drinking, itch scratching, and bed-bathing. Given our previous success with uniform sampling, we consistently employed uniform sampling for all generalized performance tests, as reflected in the experimental results shown in Fig. 4. As our algorithm relies on a preference-decoupled approach, we needed to pre-define a task reward, $\hat{r}_{task}$, to serve as prior knowledge for the reward model. In the feeding task, we define $\hat{r}_{task} = -\|d\|^2$ as in [24], where d represents the distance from the target location to the robot's end effector. Our experiments show that this definition of distance-to-target-point is also effective in other settings. From Fig. 4, we notice that across all metrics and environments, PPO with true reward consistently outperforms the other two methods, showcasing its robustness and efficiency in achieving high returns, task success rates, and preference returns. Our Decoupled PrefPPO with GR, while not as effective as the true reward method, still provides a viable alternative with moderate performance. The Decoupled PrefPPO with PC, however, demonstrates significant challenges in effectively completing the tasks, indicating that pairwise preference feedback may not be effective in optimal policy learning.

4.4 Influences of Group Ranking Size M

The parameter M plays a crucial role in determining how many trajectories are simultaneously used to request human expert rankings. We conducted tests with four different values of M, namely $M \in \{2, 5, 10, 15\}$, in the feeding task. The experimental results are shown in Table 2. The results indicate that a larger number of comparisons does not necessarily yield better results. Excessive comparisons can increase the risk of the reward model overfitting, hampering its performance. In contrast, an appropriate number of trajectories provides richer information compared to simple pairwise trajectory comparisons. Our experiments indicate that $M = 5$ represents a suitable value.

Discussion. The entire process is performed in a simulated environment, and one potential challenge to transfer to real-world applications is the ability of a human to rank the order of $M \gg 2$ different trajectories as efficiently and effectively as the case for $M = 2$ or close to 2. As indicated in [36], the human performance of feedback may not scale with M as a manually designed program described in Sect. 4.1.

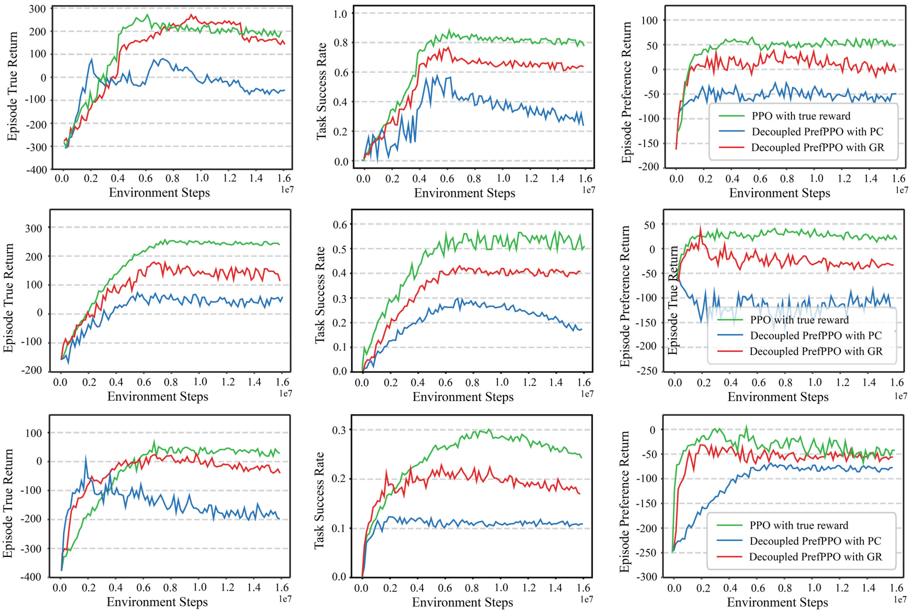


Fig. 4. The learning curves in drinking (top row), itch scratching (middle row), and bed bathing (bottom row) assistance tasks. The results of our 'Decoupled PrefPPO with GR' method are compared to 'PPO with true reward' and 'Decoupled PrefPPO with PC'.

5 Conclusions

Our aim is to create a personalized robot for various human-robot interactions. In this paper, we introduce preference-based RL with a ranking of multiple trajectories based on human preferences. We utilize a reward learned from preference-based RL to refine the robot's policy, ensuring alignment with human preferences. In addition, we integrate a sketchy reward based on prior knowledge to boost task exploration. Our experimental results confirm the superior performance of our approach in complex interactive tasks, underscoring its efficacy in facilitating personalized human-robot interaction.

References

1. Akrour, R., Schoenauer, M., Sebag, M.: April: active preference learning-based reinforcement learning. In: European Conference on Machine Learning and Knowledge Discovery in Databases, pp. 116–131. Springer (2012)
2. Archambeau, C., Caron, F.: Plackett-luce regression: a new Bayesian model for polychotomous data. In: Conference on Uncertainty in Artificial Intelligence (2012)
3. Bauer, A., Wollherr, D., Buss, M.: Human-robot collaboration: a survey. Int. J. Humanoid Rob. **5**(01), 47–66 (2008)
4. Biyik, E., Sadigh, D.: Batch active preference-based learning of reward functions. In: Conference on Robot Learning, pp. 519–528. PMLR (2018)
5. Bradley, R.A., Terry, M.E.: Rank analysis of incomplete block designs: I. the method of paired comparisons. Biometrika **39**(3/4), 324–345 (1952)
6. Brown, D., Goo, W., Nagarajan, P., Niekum, S.: Extrapolating beyond suboptimal demonstrations via inverse reinforcement learning from observations. In: International Conference on Machine Learning, pp. 783–792. PMLR (2019)
7. Christiano, P.F., Leike, J., Brown, T., Martic, M., Legg, S., Amodei, D.: Deep reinforcement learning from human preferences. In: Advances in Neural Information Processing Systems, vol. 30 (2017)
8. Clabaugh, C., Matarić, M.: Robots for the people, by the people: personalizing human-machine interaction. Sci. Robot. **3**(21), eaat7451 (2018)
9. Claure, H., et al.: Fairness and transparency in human-robot interaction. In: ACM/IEEE International Conference on Human-Robot Interaction, pp. 1244–1246. IEEE (2022)
10. El-Shamouty, M., Wu, X., Yang, S., Albus, M., Huber, M.F.: Towards safe human-robot collaboration using deep reinforcement learning. In: IEEE International Conference on Robotics and Automation, pp. 4899–4905. IEEE (2020)
11. Erickson, Z., Gangaram, V., Kapusta, A., Liu, C.K., Kemp, C.C.: Assistive gym: a physics simulation framework for assistive robotics. In: IEEE International Conference on Robotics and Automation, pp. 10169–10176. IEEE (2020)
12. Fürnkranz, J., Hüllermeier, E., Cheng, W., Park, S.H.: Preference-based reinforcement learning: a formal framework and a policy iteration algorithm. Mach. Learn. **89**, 123–156 (2012)
13. Ge, L., et al.: Axioms for AI alignment from human feedback. arXiv preprint arXiv:2405.14758 (2024)
14. Haddadin, S., Albu-Schäffer, A., Hirzinger, G.: Requirements for safe robots: measurements, analysis and new insights. Int. J. Robot. Res. **28**(11–12), 1507–1527 (2009)

15. 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)
16. Hornik, K., Stinchcombe, M., White, H.: Multilayer feedforward networks are universal approximators. Neural Netw. **2**(5), 359–366 (1989)
17. Howard, A.: Are we trusting AI too much? Examining human-robot interactions in the real world. In: Proceedings of the ACM/IEEE International Conference on Human-Robot Interaction, p. 1 (2020)
18. Ibarz, B., Leike, J., Pohlen, T., Irving, G., Legg, S., Amodei, D.: Reward learning from human preferences and demonstrations in atari. In: Advances in Neural Information Processing Systems, vol. 31 (2018)
19. Kapusta, A., et al.: Personalized collaborative plans for robot-assisted dressing via optimization and simulation. Auton. Robot. **43**, 2183–2207 (2019)
20. Khosla, P., et al.: Supervised contrastive learning. Adv. Neural. Inf. Process. Syst. **33**, 18661–18673 (2020)
21. Lee, K., Smith, L.M., Abbeel, P.: Pebble: feedback-efficient interactive reinforcement learning via relabeling experience and unsupervised pre-training. In: International Conference on Machine Learning, pp. 6152–6163. PMLR (2021)
22. Li, G., Gomez, R., Nakamura, K., He, B.: Human-centered reinforcement learning: a survey. IEEE Trans. Hum.-Mach. Syst. **49**(4), 337–349 (2019)
23. Liang, X., Shu, K., Lee, K., Abbeel, P.: Reward uncertainty for exploration in preference-based reinforcement learning. In: International Conference on Learning Representations (2021)
24. Liu, M., Chen, C.: Task decoupling in preference-based reinforcement learning for personalized human-robot interaction. In: IEEE/RSJ International Conference on Intelligent Robots and Systems, pp. 848–855. IEEE (2022)
25. Liu, M., Xiao, C., Chen, C.: Perspective-corrected spatial referring expression generation for human-robot interaction. IEEE Trans. Syst. Man Cybern. Syst. **52**(12), 7654–7666 (2022)
26. Luce, R.D.: Individual Choice Behavior, vol. 4. Wiley, New York (1959)
27. Maystre, L., Grossglauser, M.: Fast and accurate inference of plackett–luce models. In: Advances in Neural Information Processing Systems, vol. 28 (2015)
28. Mnih, V., et al.: Human-level control through deep reinforcement learning. Nature **518**(7540), 529–533 (2015)
29. Munzer, T., Toussaint, M., Lopes, M.: Preference learning on the execution of collaborative human-robot tasks. In: 2017 IEEE International Conference on Robotics and Automation, pp. 879–885. IEEE (2017)
30. Myers, V., Biyik, E., Anari, N., Sadigh, D.: Learning multimodal rewards from rankings. In: Conference on Robot Learning, pp. 342–352. PMLR (2022)
31. Obaigbena, A., Lottu, O.A., Ugwuanyi, E.D., Jacks, B.S., Sodiya, E.O., Daraojimba, O.D.: Ai and human-robot interaction: a review of recent advances and challenges. GSC Adv. Res. Rev. **18**(2), 321–330 (2024)
32. Oliff, H., Liu, Y., Kumar, M., Williams, M., Ryan, M.: Reinforcement learning for facilitating human-robot-interaction in manufacturing. J. Manuf. Syst. **56**, 326–340 (2020)
33. Palan, M., Shevchuk, G., Charles Landolfi, N., Sadigh, D.: Learning reward functions by integrating human demonstrations and preferences. In: Robotics: Science and Systems (2019)
34. Park, J., Seo, Y., Shin, J., Lee, H., Abbeel, P., Lee, K.: Surf: semi-supervised reward learning with data augmentation for feedback-efficient preference-based reinforcement learning. In: International Conference on Learning Representations (2021)

35. Pleskac, T.J.: Decision and choice: Luce's choice axiom. Int. Encycl. Soc. Behav. Sci. **5**, 895–900 (2015)
36. Sankaran, S., Derechin, J., Christakis, N.A.: Curmelo: the theory and practice of a forced-choice approach to producing preference rankings. PLoS ONE **16**(5), e0252145 (2021)
37. Schulman, J., Wolski, F., Dhariwal, P., Radford, A., Klimov, O.: Proximal policy optimization algorithms. arXiv preprint arXiv:1707.06347 (2017)
38. Silver, D., et al.: Mastering the game of go with deep neural networks and tree search. Nature **529**(7587), 484–489 (2016)
39. Song, F., et al.: Preference ranking optimization for human alignment. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 38, pp. 18990–18998 (2024)
40. Stiennon, N., et al.: Learning to summarize with human feedback. Adv. Neural. Inf. Process. Syst. **33**, 3008–3021 (2020)
41. Sutton, R.S., Barto, A.G.: Reinforcement Learning: An Introduction. MIT Press (2018)
42. Tabrez, A., Hayes, B.: Improving human-robot interaction through explainable reinforcement learning. In: ACM/IEEE International Conference on Human-Robot Interaction, pp. 751–753. IEEE (2019)
43. Tian, L., Oviatt, S.: A taxonomy of social errors in human-robot interaction. ACM Trans. Hum.-Robot Interact. **10**(2), 1–32 (2021)
44. Tien, J., Brown, D.: Causal confusion and reward misidentification in preference-based reward learning. In: International Conference on Learning Representations (2023)
45. Vasconez, J.P., Kantor, G.A., Cheein, F.A.A.: Human-robot interaction in agriculture: a survey and current challenges. Biosys. Eng. **179**, 35–48 (2019)
46. Wilson, A., Fern, A., Tadepalli, P.: A Bayesian approach for policy learning from trajectory preference queries. In: Advances in Neural Information Processing Systems, vol. 25 (2012)
47. Wirth, C., Akrour, R., Neumann, G., Fürnkranz, J., et al.: A survey of preference-based reinforcement learning methods. J. Mach. Learn. Res. **18**(136), 1–46 (2017)
48. Woodworth, B., Ferrari, F., Zosa, T.E., Riek, L.D.: Preference learning in assistive robotics: observational repeated inverse reinforcement learning. In: Machine Learning for Healthcare Conference, pp. 420–439. PMLR (2018)
49. Zhan, H., Tao, F., Cao, Y.: Human-guided robot behavior learning: a GAN-assisted preference-based reinforcement learning approach. IEEE Robot. Autom. Lett. **6**(2), 3545–3552 (2021)
50. Zhang, C., Chen, J., Li, J., Peng, Y., Mao, Z.: Large language models for human-robot interaction: a review. Biomimetic Intell. Robot. 100131 (2023)
51. Zhang, R., Lv, Q., Li, J., Bao, J., Liu, T., Liu, S.: A reinforcement learning method for human-robot collaboration in assembly tasks. Robot. Comput.-Integr. Manuf. **73**, 102227 (2022)

Synthesizing Operationally Safe Controllers for Human-in-the-Loop Human-in-the-Plant Hybrid Close Loop Systems

Ayan Banerjee[1](✉), Imane Lamrani[1,2], and Sandeep K. S. Gupta[1]

[1] IMPACT Lab, Arizona State University, Tempe, USA
{abanerj3,imane.lamrani,sandeep.gupta}@asu.edu,
imane.lamrani@nikolamotor.com
[2] Nikola Motors, Phoenix, USA

Abstract. Human inputs are considered external disturbances in traditional certified safe controller synthesis approaches and are modeled using non-causal random variables with an assumed parameterized distribution. However, (human) safety-critical autonomous systems such as medical devices and autonomous cars operate in hybrid closed loop (HCL) mode, where humans are required to either provide control inputs, perturb the physical system being controlled (called a plant in control theory), or completely override the autonomous system (e.g. in Level 3 autonomy). Hence, the system often is subjected to causal human actions in operational deployment, that cannot be accurately modeled using non-causal distributions - leading to "flawed" safety-certified designs susceptible to operational failures in presence of unmodeled human actions (e.g. Boeing 747 Max MCAS failure). We propose a human-in-the-loop (HIL)-human-in-the-plant (HIP) approach towards synthesizing controllers for safety-critical autonomous systems where the human mind (HIL), the human body (HIP) and the real world controller (RWC) are modeled as an unified system. A three-way interaction is considered: a) through personalized inputs and biological feedback processes between HIP and HIL, b) through sensors and actuators between RWC and HIP, and c) through personalized configuration changes and data feedback between HIL and RWC. We extend the control Lyapunov theory by generating barrier function (CLBF) under human action plans, model the HIL as a combination of a Markov Chain (MC) for spontaneous events and a Fuzzy inference system (FIS) for event responses, the RWC as a black box, and integrate the HIL-HIP model with neural architectures that can learn CLBF certificates. Our main result is Theorem 1, which shows that if human actions are in the p-domain of attraction of the MC-FIS model of HIL, the synthesized controller satisfies safety properties (specified in Symbol Temporal Logic (STL)) with probability at least p. We demonstrate the capability of safe controller synthesis of our approach on two HCL applications: a) autonomous vehicle braking system, and b) automate insulin delivery for Type 1 Diabetes.

This work is partially funded by DARPA AMP-N6600120C4020, and FIRE - P000050426 and NSF FDT-BioTech 2436801.

A. Antonacopoulos et al. (Eds.): ICPR 2024, LNCS 15329, pp. 17–35, 2025.
https://doi.org/10.1007/978-3-031-78110-0_2

Keywords: Controller synthesis integrated with human · human-in-the-loop · control Lyapunov functions

1 Introduction

Safety-criticality implies that the operation of the autonomous system (AS) can cause harm to the human participants who are affected by the AS goal [2,6]. Given the impending risks to the human user, safety-critical applications most typically operate in a hybrid closed loop (HCL) mode [3]. Here, a human-in-the-loop (HIL) is in charge of starting and stopping automation and can provide manual inputs whenever the user perceives safety risks or operational inefficiency. HCL applications are observed in level 3 autonomy as seen in medical devices [7] (automated insulin delivery, AID) or autonomous vehicles (AV) [23]. HCL operation in such human-centered AS results in a HIL with human-in-the-plant (HIL-HIP) system model (Sect. 3), where the human user is the monitor/decision maker and also part of the physical plant controlled by the AS [17,18] (Fig. 1).

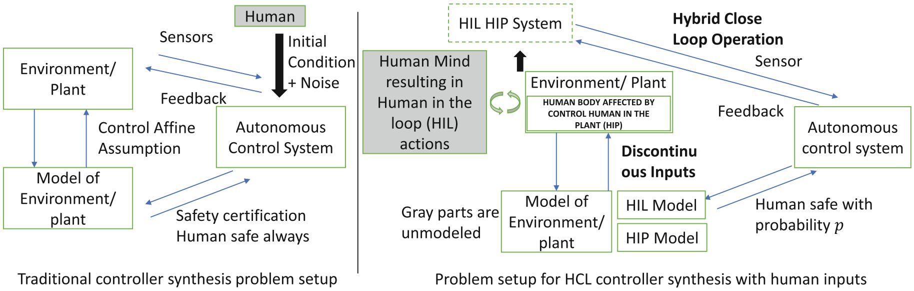

Human
Sensors
Initial Condition + Noise
Environment/ Plant
Feedback
Control Affine Assumption
Autonomous Control System
Model of Environment/ plant
Safety certification Human safe always
Traditional controller synthesis problem setup
HIL HIP System
Hybrid Close Loop Operation
Sensor
Human Mind resulting in Human in the loop (HIL) actions
Environment/ Plant
HUMAN BODY AFFECTED BY CONTROL HUMAN IN THE PLANT (HIP)
Feedback
Autonomous control system
Discontinuous Inputs
Gray parts are unmodeled
Model of Environment/ plant
HIL Model
HIP Model
Human safe with probability p
Problem setup for HCL controller synthesis with human inputs

Fig. 1. Traditional safety assured controller synthesis model human outside the system as noise input with known distribution. Hybrid close loop systems may have causal human inputs that may not fit a non-causal distribution.

In HCL, often the human user is physically part of the plant and a control action affecting the plant also affects the human body (Fig. 1, second panel). For example, if an AV accelerates or brakes too fast, the human body experiences its effects. Moreover, in case of crashes the human body bears significant risks. In case of AID systems, the human body is directly affected by the drug infusions decided by the controller. Hence, in HCL systems, the human body is a part of the plant (HIP) and its physiological responses affects the AS operation. On the other hand, HCL operation, as mandated by many safety certification agencies such as US Food and Drug Administration (FDA) or Federal Aviation Administration (FDA) [7,20], implies that the human (mind) should be able to decide on interventions to the control actions. In case of AID, the human can decide to take food, and stop or provide additional infusion. In case of AV, the human user can provide additional braking action on perceiving impending crash.

Thus in HCL system, the human mind affects the control actions by intervening with the AS through actuation termed as HIL actions. The HIL actions are in response to the experiences of the HIP due to the operation of the AS, and hence are causal actions [3,4]. Such human integrated operation in HCL is not modeled in the traditional workflow of safe controller synthesis for an AS (Fig. 1 first panel).

Existing safe controller synthesis process assume a control affine system model, where the plant state X is assumed to follow the dynamics in Eq. 1.

$$\dot{X} = f(X) + g(X)\pi(X), \tag{1}$$

where $f(.)$ is the un-perturbed plant response model, $g(.)$ is the input effect, and $\pi(.)$ is a controller that computes an input to the plant based on the plant state X. In HCL, the input to the plant is given by: $u = \pi(X) + u_{ex}$, where $u_{ex} \in U_{ex}$ is an external input from the human user. Despite human user being an integral part of the AS operation, controller synthesis using the control affine assumption consider the human as external to the system (Fig. 1).

The control affine assumption enables key mathematical advantages in the controller synthesis problem: a) the control action is assumed to be Lipschitz continuous and can be obtained by solving a continuous nonlinear optimization problem, and b) candidate Lyapunov function for stability of Eq. 1 is the sum of squares of the state variables in X. If the Lipschitz assumption of inputs is broken then the sum of squares of state variable X is no longer a candidate Lyapunov function and hence solving the nonlinear controller synthesis problem becomes very difficult and to the best of our knowledge no current solution exists.

Traditional nonlinear controller synthesis techniques are applied for HCL systems by assuming a fixed distribution of human inputs. Any user that satisfies the distribution is termed as an *average user*. The representative value of the distribution such as the mean is used to obtain a deterministic solution, and subsequently the moments of the distribution such as standard deviation is used to provide stochastic guarantees. This approach allows traditional approaches to derive an "optimal" solution to the controller synthesis problem since non-causal standalone human inputs only act as initial or boundary conditions.

Large scale deployment and day-to-day usage imply that a significant number of users will be non-conformal to the "average user" resulting in novel usage scenarios. To replicate the performance obtained in the certification process, the real user may undertake *personalization actions*, which are sequences of external inputs or system configuration changes applied with/without expert advisory agent (clinicians) consultations [5,15]. Such input sequences may have a causal relation with the HIP state

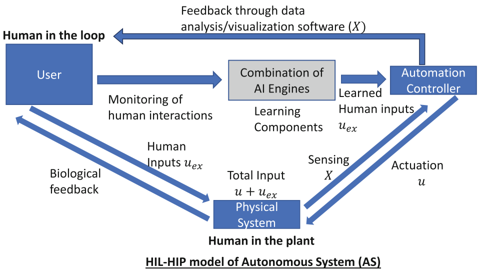


Fig. 2. HIL-HIP autonomous systems (AS).

X, are out of distribution, and may invalidate safety certificates of the synthesized controller. Such unverified personalizations can jeopardize operational safety.

Contributions: In this paper, we assume that the real-world controller (RWC) or $\pi(.)$ in AS is safety certified with the control affine assumption, for the "average user" and a black box, and present a technique to synthesize controllers integrated with HIL-HIP actions that are safe for the human-in-the-plant. The plant model is:

$$\dot{X} = f(X) + g(X)(\pi(X) + u_{ex}), \tag{2}$$

and $u_{ex} \in U_{ex}$ is a set of personalized inputs for a given real life user. The HIL actions are modeled using a combination of a Markov Chain (MC), to model events that trigger the human actions, and a Fuzzy inference system (FIS), to model the human actions taken for a trigger event. We solve the safe controller synthesis problem with HIL-HIP integration using neural control Lyapunov barrier function (CLBF) strategy [11].

Theorem 1 establishes that controllers designed following the proposed HIL-HIP strategy are safe with a safety tuning probability (STP) p if the human actions are within the p domain of attraction (DoA) of the underlying MC model. This technique is validated by synthesizing certified safe controllers for automated insulin delivery (AID) to control glucose levels in Type 1 Diabetes (T1D) and autonomous braking systems (ABS) in vehicles.

2 Motivating Examples

HCL operating mode is seen in two exemplary autonomous systems: the autonomous braking (ABS) and the automated insulin delivery (AID) systems.

2.1 Autonomous Braking (ABS) System

Assume that the car A in Fig. 3 is braking autonomously, while keeping a safe distance d between cars A and B. The controller assumes a control affine model of the vehicle kinetics given by Eq. 3.

$$\begin{gathered} \dot{a}_A = -0.01 s_{AB} + 0.737 - 0.3(v_{AB}) - 0.5 a_A, \\ \dot{v}_{AB} = 0.1 a_A + u_A,\ \dot{s}_{AB} = -v_{AB} - 2.5. \end{gathered} \tag{3}$$

$v_{AB} = v_A - v_B$ is the relative velocity of car A with respect to car B, s_{AB} is the distance, and u_A is the deceleration input from the ABS and the human.

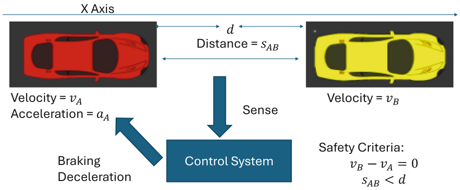


Fig. 3. Autonomous Braking System.

While the autonomous system can apply the brakes and stop the car before a major accident, a human driver can override and provide manual braking inputs. Hence, $u_A = \pi(a_A, v_B - v_A, s_{AB}) + u_h$, where u_h is the manual human braking input and $\pi(.)$ is the control law. At this point, the ABS is still active and can provide additional braking or anti-lock braking system facilities by countering the human braking input. The aim of the anti-lock braking systems is to prevent skid. Skid condition is defined as an upper limit on the ratio of the deceleration to the velocity, i.e. $a_A/v_A < \tau_{th}$. The aim of the controller is to satisfy if $s_{AB} < d$ then $v_{AB} = 0$ and the skid condition. The human input u_h depends on two factors not modeled in the traditional controller synthesis approaches with control affine assumption:

a) reaction time of the human driver: the human driver can manually engage braking after the user realizes that the car A is too close to car B, i.e. $s_{AB} < d_r$. However, the driver has a reaction time τ_r, where the driver engages manual braking τ_r time after $s_{AB} = d_r$.

b) driving behavior of the driver: when the human driver manually engages braking, the braking force u_h varies based on several unmodeled factors.

2.2 Automated Insulin Delivery (AID) System

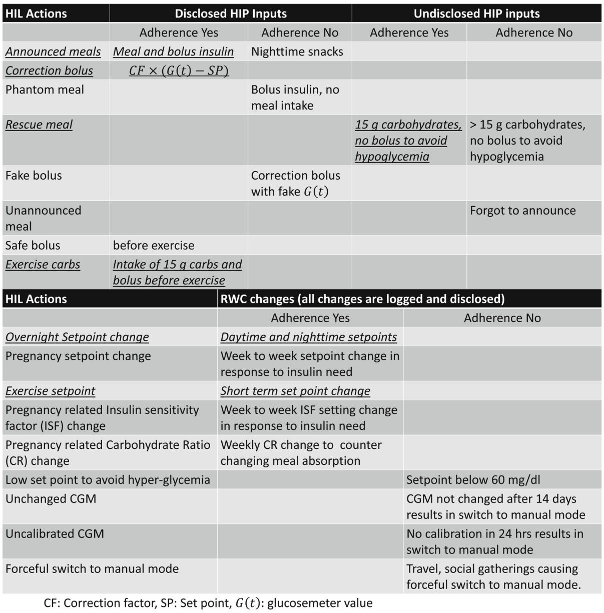
The AID system senses blood glucose using continuous glucose monitors (CGM) and automatically delivers insulin to keep CGM between a lower limit of 70 mg/dL and higher limit of 180 mg/dl. In the AID system, the glucose insulin dynamics is given by the Bergman Minimal Model (BMM) represented as:

$$\dot{\delta i}(t) = -n\delta i(t) + p_4 u_1(t), \tag{4}$$

$$\dot{\delta i}_s(t) = -p_1 \delta i_s(t) + p_2(\delta i(t) - i_b), and \tag{5}$$

$$\dot{\delta G}(t) = -\delta i_s(t) G_b - p3(\delta G(t)) + u2(t)/VoI. \tag{6}$$

The input vector consists of the overnight basal insulin level i_b and the glucose appearance rate in the body u_2. The output vector is comprised of the blood insulin level i, the interstitial insulin level i_s, and the blood glucose level G. The parameters p_1, p_2, p_3, p_4, n, and $1/V_oI$ are all patient specific coefficients. The controller decides on microbolus inputs every 5 min. In addition, users can also manually provide priming bolus Θ_u to prepare for an unplanned glycemic event such as a meal. A complete list of manual intervention in the HCL operating model of the AID is shown in Fig. 4.



HIL Actions	Disclosed HIP Inputs		Undisclosed HIP inputs	
	Adherence Yes	Adherence No	Adherence Yes	Adherence No
Announced meals	*Meal and bolus insulin*	Nighttime snacks		
Correction bolus	$CF \times (G(t) - SP)$			
Phantom meal		Bolus insulin, no meal intake		
Rescue meal			*15 g carbohydrates, no bolus to avoid hypoglycemia*	> 15 g carbohydrates, no bolus to avoid hypoglycemia
Fake bolus		Correction bolus with fake $G(t)$		
Unannounced meal				Forgot to announce
Safe bolus	before exercise			
Exercise carbs	*Intake of 15 g carbs and bolus before exercise*			

HIL Actions	RWC changes (all changes are logged and disclosed)	
	Adherence Yes	Adherence No
Overnight Setpoint change	*Daytime and nighttime setpoints*	
Pregnancy setpoint change	Week to week setpoint change in response to insulin need	
Exercise setpoint	*Short term set point change*	
Pregnancy related Insulin sensitivity factor (ISF) change	Week to week ISF setting change in response to insulin need	
Pregnancy related Carbohydrate Ratio (CR) change	Weekly CR change to counter changing meal absorption	
Low set point to avoid hyper-glycemia		Setpoint below 60 mg/dl
Unchanged CGM		CGM not changed after 14 days results in switch to manual mode
Uncalibrated CGM		No calibration in 24 hrs results in switch to manual mode
Forceful switch to manual mode		Travel, social gatherings causing forceful switch to manual mode.

CF: Correction factor, SP: Set point, $G(t)$: glucosemeter value

Fig. 4. Categorization of HIL actions for AID system into Disclosed and Undisclosed inputs as well as RWC changes compiled from [8]. For each type whether they are adherent or non-adherent to AID operational guidance. Demonstrates interplay between RWC, physiological and mental state of HIL-HIP in the context of maintaining operational safety in a hybrid closed loop system.

3 System Model and Preliminaries

A **plant**, e.g. an autonomous car with human driver or a human body with AID and an on-body CGM, is described by the N dimensional real state vector $X \in \mathcal{X}$, where $\mathcal{X} \subset \mathcal{R}^N$ is the state space of the plant. For AID, X is a 3×1 vector, with CGM, interstitial insulin, and plasma insulin as elements.

An **autonomous real world controller** (RWC) $\pi(X)$ uses sensors (CGM for AID) on the plant (human body) to monitor its current state, and actuators (insulin pump) to deliver control inputs $u \in \mathcal{R}$ (micro bolus insulin). The control task of the RWC is to drive a state variable $x_i \in X$ (say CGM) to a set point x_i^g (say 120 mg/dL).

In **control affine systems**, the response of the plant to control inputs is modeled as a linear combination of the unperturbed continuous time state evolution of the plant, $f(X)$ and control input effect, $g(X)$ for the input u (Eq. 1). For T1D the BMM [24] expresses $f(X)$ and $g(X)$ in the form of a set of nonlinear differential equations. The control inputs $u = \pi(X)$ is a function of the sensed state variables. Human inputs are external inputs U_{ex} in addition to the control inputs u (Fig. 4).

HIL as a Controller: Human users play an integral part in AID usage (ControlIQ [8]), where they can provide spontaneous inputs (such as meal) to the HIP or take part in critical hazard mitigation (rescue meal to mitigate hypoglycemia). The actions can be categorized into (Fig. 4): a) disclosed or undisclosed HIP inputs, b) adherence to clinician guidance, and c) changes to RWC configuration. The underlined and italicized text denotes the profile of the average user for which the RWC is safety certified. There are several HIL actions that may adhere to clinician guidelines but are not commensurate with the average user profile, and are not certified safe.

The HIL-HIP **AS** (Fig. 2) consists of AI engines that learn from human interaction and give personalization plans, which is a temporally aligned finite sequence of control tasks interleaved with external inputs $(x_i^g(t_1)x_i^g(t_2)\ u_{ex}^1 x_i^g(t_3)\ u_{ex}^2\ x_i^g(t_4)\dots)$ driven by actions in Fig. 4. The **safety** is defined on subsets of $\mathcal{X}$, using Signal Temporal Logic (STL) ϕ [12]. Eventual safety indicates that $\exists \tau \in \mathcal{R} : \phi \models \top\ \forall t > \tau$, i.e. ϕ will be true after some time τ.

Safety certificate for an RWC $\pi(X)$ is the existence of a given subset $C \subset \mathcal{X}$, with forward invariance ($\mathbb{FI}$) property [11], which states that if the initial state of the AS is in C then the closed loop system dynamics in Eq. 1 keeps the initial state within C in absence of any external perturbation. A set C has $\mathbb{FI}$ property if there exists a control Lyapunov function $V(X), \forall X \in C$ such that $\forall X \in C$, $V(X) > 0$, $V(X^g) = 0$, for the set point X^g, and Eq. 7 holds true.

$$\forall X \in C, \exists \lambda > 0 : L_f V(X) + L_g V(X)\pi(X) + \lambda V(X) < 0, \tag{7}$$

where L_f and L_g are Lie derivatives of $V(X)$ along the direction of $f(X)$ and $g(X)$, respectively.

Operational Safety: An AS has operational safety if the safety STL satisfied by the AS model is also satisfied in real-world deployments.

3.1 Formal Problem Statement

Given:

- a RWC, $\pi_{nom}(X)$
- a set $C \in \mathcal{X}$ such that $\forall X \in C$ at $t = 0$, the trajectory $X(t) \in C | \dot{X(t)} = f(X(t)) + g(X(t))\pi_{nom}(X(t)), t > 0$,
- a safety tuning probability (STP) value p
- a set of personalized inputs $U_{ex} = \{u_{ex}\}$.

Find: a RWC $\pi_{NN}(X)$ such that $\forall X \in C$, for any $t > 0$

$$P(X(t) \in C) > p | \dot{X}(t) = f(X(t)) + g(X(t))(\pi_{NN}(X(t)) + u_{ex}), \tag{8}$$

Significance of HIL-HIP Architecture: In the HIL-HIP architecture the safe design is parameterized by STP p, which implies that the design is safe for human actions whose probability of occurrence is p based on a human action model. A large value of p indicates that the AS is safe for most common actions, whereas a smaller value of p requires safe operation for even unusual actions. CLBF is easier to find with larger p values (e.g. p = 0.95 in the AID example in Sect. 5) however, with a smaller p CLBF may not be found and hence the AS may not be certified safe. The STP can be used by a real user and their advisors to determine if their action model is compliant with the safety certificate. Traditional control system designs do not provide such safe personalization hook during operation.

4 Solution and Proof of Safety

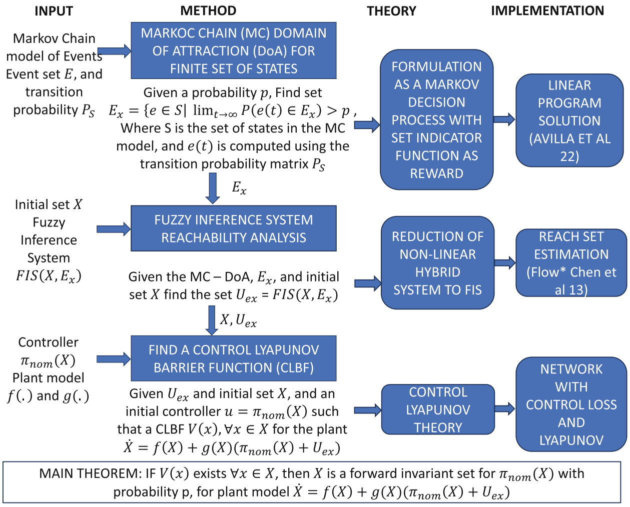

INPUT
METHOD
THEORY
IMPLEMENTATION
Markov Chain model of Events Event set E, and transition probability P_S
MARKOC CHAIN (MC) DOMAIN OF ATTRACTION (DoA) FOR FINITE SET OF STATES
Given a probability p, Find set $E_x = \{e \in S \mid \lim_{t\to\infty} P(e(t) \in E_x) > p$, Where S is the set of states in the MC model, and $e(t)$ is computed using the transition probability matrix P_S
FORMULATION AS A MARKOV DECISION PROCESS WITH SET INDICATOR FUNCTION AS REWARD
LINEAR PROGRAM SOLUTION (AVILLA ET AL 22)
E_x
Initial set X Fuzzy Inference System $FIS(X, E_x)$
FUZZY INFERENCE SYSTEM REACHABILITY ANALYSIS
Given the MC – DoA, E_x, and initial set X find the set $U_{ex} = FIS(X, E_x)$
REDUCTION OF NON-LINEAR HYBRID SYSTEM TO FIS
REACH SET ESTIMATION (Flow* Chen et al 13)
X, U_{ex}
Controller $\pi_{nom}(X)$ Plant model $f(.)$ and $g(.)$
FIND A CONTROL LYAPUNOV BARRIER FUNCTION (CLBF)
Given U_{ex} and initial set X, and an initial controller $u = \pi_{nom}(X)$ such that a CLBF $V(x), \forall x \in X$ for the plant $\dot{X} = f(X) + g(X)(\pi_{nom}(X) + U_{ex})$
CONTROL LYAPUNOV THEORY
NETWORK WITH CONTROL LOSS AND LYAPUNOV
MAIN THEOREM: IF $V(x)$ exists $\forall x \in X$, then X is a forward invariant set for $\pi_{nom}(X)$ with probability p, for plant model $\dot{X} = f(X) + g(X)(\pi_{nom}(X) + U_{ex})$

Fig. 5. Solution: Model HIL as a combination of MC and FIS. Derive action sequences that are in p-DoA of MC. Synthesize controller using neural CLBF under actions in p-DoA. CLBF controller is safe with probability p (Theorem 1).

To address safety of an AS under personalized human actions, the AS is modeled as a HIL-HIP unified system, with three way interactions (Fig. 2) between the HIL which is modeled as a joint Markov Chain (MC) for spontaneous events and fuzzy inference system (FIS) for human responses to the events, RWC and the HIP. The HIL controller receives information from the RWC through data analysis & visualization app and the natural biological feedback mechanism from the HIP. Based on advice from external advisory agents (clinicians), the HIL controller decides: a) inputs to the HIP, e.g., meal or bolus insulin, and b) inputs to the RWC, e.g., settings change. Safe HIL-HIP controller synthesis has three step (Fig. 5):

Step 1: Find the domain of attraction (DoA) of MC model E_x with minimum probability p. Starting from an initial set of states E_I the DoA is the set of MC states E_x that will occur at some point of time with at least p probability of occurrence. The MC extends a Markov Decision Process (MDP) with reward function same as the indicator function for the set E_x [25]. Solution of a linear program for value function maximization gives the DoA [25].

Step 2: For the DoA E_x, find the reach set U_{ex} of the FIS model for an initial set X of states. We show that a hybrid system can be reduced to a FIS model.

The reach set estimation method of the hybrid system model of the FIS gives an over-approximation of the reach set U_{ex}.

Step 3: For the U_{ex}, search a CLBF $V(X)$ using the neural architecture with control loss and Lypunov loss to derive π_{NN} with the plant model of Eq. 2.

To develop the HIL-HIP system, data driven learning is performed in two stages (Fig. 6): **Stage 1**, the MC and FIS are learned for human action, which is fed to **Stage 2**, the neural CLBF architecture that learns the RWC.

4.1 Modeling External Events with Markov Chain

The MC model is used to capture the spontaneous events that occur during long term usage of the AS. MC model states denote the unique events that are triggered due to the day to day activities of the HIP. Each unique event is parsed from the usage data of the AS. The set of unique events is denoted by $\mathbf{S}$. The transition from state $s_i \in \mathbf{S}$ to s_j is tabulated from the usage data. Transition probabilities are computed using a conditional probability computation method and Bayes theorem. For each event s_i, we search the set of other events $\mathbf{S}_i = \{s_j\}$ such that s_j is the next state after s_i. For each s_j we count the number of times n_{ij} that event s_j occurred after event s_i. The transition probability $P_S(i,j)$, for transiting from event s_i to s_j is computed as $P_S(i,j) = n_{ij}/\sum_{\forall s_j \in \mathbf{S}_i} n_{ij}$.

4.2 Finding DoA of MC with Minimum Probability p

A MDP with the state space as S and the reward function as the indicator function $\mathbb{1}_{E_x}$ is considered, where E_x is the DoA. The value function is Eq. 9,

$$v^{pol}(e) := \limsup_{M \to \infty} \frac{1}{M} E_e^{\pi} \Big[\sum_{t=0}^{M-1} \mathbb{1}_{E_x}(S) \Big], \tag{9}$$

where E_e^{π} is the expected value function given as $P_S(e_j|e)$, where e_j is any next state and $v^{pol}(e)$ denotes the value function for a given policy *pol* in state e.

Lemma 1. *The set of states explored by the optimal policy with value function $v^*(e) > p$, gives the reach set of the MC (S, P_S) starting from state e, where v^* is given by Eq. 9.*

The lemma is derived from Theorem III.4 in [25]. The DoA E_x of the MC (S, P_S) with minimum probability p is obtained by solving the linear program RealP in [25].

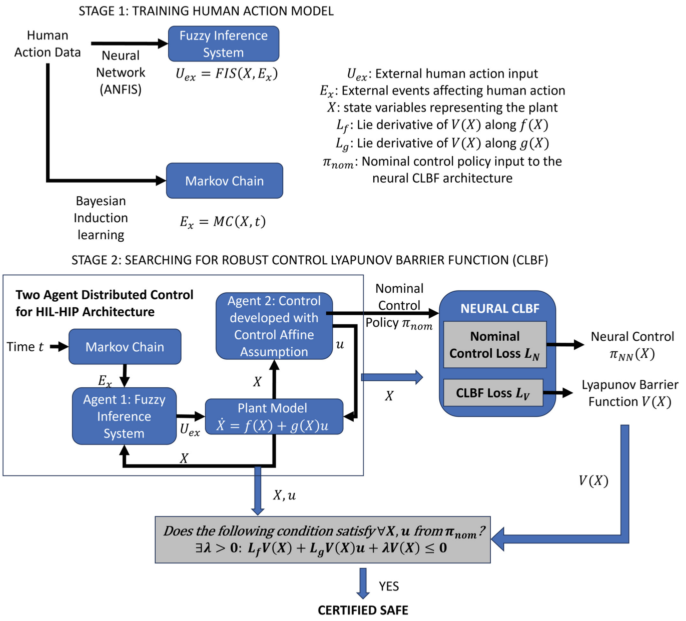
Computational Complexity: The solution to the linear program has a complexity of $O(N^3 log(N/\delta))$, where δ is the error tolerance [10].

4.3 Learning a FIS from Data

The human action database $\mathcal{D} = \{E_x, X(t), U_{ex}\}$ consists of traces $X(t)$ of the state variables X over time t, external events E_x that either affect the AS controller configuration or affect the HIP component, and the human action, U_{ex}, taken by the human in response to observation of the state variables and external events. We model the human action as a function of X, E_x using techniques such as FIS. The membership functions and fuzzy rules is learned from data using techniques such as adaptive network-based fuzzy inference system (ANFIS) [22]. The output is human action model $U_{ex} = FIS(X, E_x)$.

4.4 Finding Reach Set for FIS

Lemma 2. *For every FIS output u_{ex} for a given X and event E_x, there exists an execution of a rectangular hybrid system $(Q, V, \mathcal{F} : 2^V \rightarrow \mathcal{R}^{|V|}, Inv)$ that provides the final output u_{ex} of one of its continuous states $v \in V$.*



STAGE 1: TRAINING HUMAN ACTION MODEL
Human Action Data
Neural Network (ANFIS)
Fuzzy Inference System
$U_{ex} = FIS(X, E_x)$
U_{ex}: External human action input
E_x: External events affecting human action
X: state variables representing the plant
L_f: Lie derivative of $V(X)$ along $f(X)$
L_g: Lie derivative of $V(X)$ along $g(X)$
π_{nom}: Nominal control policy input to the neural CLBF architecture
Markov Chain
Bayesian Induction learning
$E_x = MC(X, t)$
STAGE 2: SEARCHING FOR ROBUST CONTROL LYAPUNOV BARRIER FUNCTION (CLBF)
Two Agent Distributed Control for HIL-HIP Architecture
Time t
Markov Chain
E_x
Agent 1: Fuzzy Inference System
U_{ex}
Agent 2: Control developed with Control Affine Assumption
u
X
Plant Model $\dot{X} = f(X) + g(X)u$
X
Nominal Control Policy π_{nom}
X
NEURAL CLBF
Nominal Control Loss L_N
CLBF Loss L_V
Neural Control $\pi_{NN}(X)$
Lyapunov Barrier Function $V(X)$
$V(X)$
X, u
Does the following condition satisfy $\forall X, u$ from π_{nom}?
$\exists \lambda > 0$: $L_f V(X) + L_g V(X)u + \lambda V(X) \leq 0$
YES
CERTIFIED SAFE

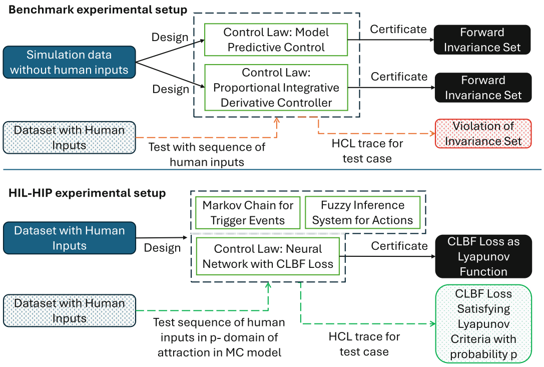
Fig. 6. Solution method for deriving safety certificate for HIL=HIP systems under the learned human action model

We prove this lemma by construction. The variable set V of the hybrid system is the same as the state vector X of the FIS. A rule in the ANFIS is of the form "IF $x_1(k) \in A_{11} \wedge \ldots \wedge x_N(k) \in A_{1N}$ THEN $u_{ex} \in B_1$", where $A_{i,j}$ is the membership function of the i^{th} rule for the j^{th} element of the state vector and B_i is the membership function of the output for the i^{th} rule. Each rule R_i learned by ANFIS is modeled as a state of the hybrid system in the set Q. The flow equation $f \in \mathcal{F}$ of each state R_i is $defuzzify(min_{j=1 \rightarrow N}(A_{i,j}))$. Defuzzification is done using the centroid mechanism. Each $A_{i,j}$ is a nonlinear sigmoid function resulting in a nonlinear continuous flow function on the power set of variables $V = X$ to the $|V| = N$ dimensional real space. The state transition condition Inv in the hybrid system is instantaneous and occurs by default resulting in a rectangular timed automata. By construction, an execution of this nonlinear hybrid system follows the exact computational steps taken by FIS to arrive at an output. Hence, the reach set of the hybrid system provides the reach set U_{ex} of the FIS. In this paper, we use the $Flow*$ [9] technique to obtain reach set of the derived hybrid system, which gives guaranteed over-approximation of U_{ex}.

Computational Complexity: The worst case computational complexity is $O(KN^2)$, where reach set is computed for K time steps ahead [9].

4.5 Safety Certificate Generation Method

The HIL-HIP controller is the integration between the human action controller ($U_{ex} = FIS(X, E_x)$, where $E_x = MC(t)$) and the AS controller $\pi(X)$. The total output of the HIL-HIP system π_{nom} is $u = \pi_{nom}(X) = U_{ex} \bigcap \pi(X)$. For this controller, given a subset $c \subset \mathcal{X}$, we want to ensure the $\mathbb{FI}$ property. This is done by showing the existence of a CLBF $V(X)$ defined for any $X \in C$. We use neural architectures with modified loss functions discussed in [11] for this purpose. The main idea is to use the neural architecture as approximators for: a) CLBF $V(X)$, and b) a neural controller π_{NN} that satisfies the Lyapunov condition for stability and safety (Eq. 7). The loss function consists of two parts:



Benchmark experimental setup
Simulation data without human inputs
Design
Design
Control Law: Model Predictive Control
Control Law: Proportional Integrative Derivative Controller
Certificate
Certificate
Forward Invariance Set
Forward Invariance Set
Dataset with Human Inputs
Test with sequence of human inputs
HCL trace for test case
Violation of Invariance Set
HIL-HIP experimental setup
Markov Chain for Trigger Events
Fuzzy Inference System for Actions
Dataset with Human Inputs
Design
Control Law: Neural Network with CLBF Loss
Certificate
CLBF Loss as Lyapunov Function
Dataset with Human Inputs
Test sequence of human inputs in p- domain of attraction in MC model
HCL trace for test case
CLBF Loss Satisfying Lyapunov Criteria with probability p

Fig. 7. Experimental design and baseline strategies. Safety certificate for traditional controller synthesis is forward invariant set. For HIL-HIP controller, the neural network with CLBF loss is the control law and the loss is the Lyapunov function.

a) **CLBF loss**, which ensures that the CLBF estimate by the neural structure $V(x)$ satisfies the relaxed condition of $V(x) < \epsilon > 0$, a small quantity.

b) **control loss**, which captures the difference in control actions by the neural controller π_{NN} and the RWC π_{nom}.

The neural structure is trained with a set of state variable and control action pairs from the RWC π_{nom}. The output of the training phase is: a) decision whether a CLBF exists or not, and b) if a CLBF exists then the trained neural architecture that gives both π_{NN} and $V(X), \forall X \in C$, According to the CLBF theory [11], if a $V(X)$ exists, then π_{NN} is one of potentially many controllers (denoted by the set of controllers $\mathcal{K}(X)$) that are safe and the neural structure that gives $V(X)$ and the corresponding forward invariant set C is a safety certificate. To ascertain whether $\pi_{nom} \in \mathcal{K}(X)$ we evaluate the CLBF condition in Eq. 7 with π_{nom} as the RWC.

Theorem 1. *If $V(X)$ exists $\forall X \in C$, then C is a forward invariant set for $\pi_{NN}(X)$ with probability p, for the plant model in Eq. 2 and if condition in Eq. 7 satisfies, then C is a forward invariant set for $\pi_{nom}(X)$ with probability p.*

Lemma 1 provides event set with occurrence probability $> p$. Lemma 2 shows that $Flow*$ reachability analysis will provide an U_{ex} that encompasses all FIS outputs that are p probable due to the over-approximation property. Existence of the Lyapunov function through the neural structure guarantees that π_{NN} will result in a safe plant when combined with U_{ex}. Hence, set C will be forward invariant for π_{NN}.

5 Evaluation

We compare our proposed HIL-HIP strategy with baseline strategies in AID and AV examples. The baseline for AID is Model predictive control (MPC) and Proportional Integrative and Derivative (PID) control, while that in AV is Linear Quadratic Gaussian (LQG) control (Fig. 7). In the traditional technique, the control law is developed assuming static human inputs, and a certificate is derived in the form of forward invariant set. However in deployment, sequence of inputs are observed that leads to violation of the invariant set.

On the other hand, for HIL-HIP architecture, the sequence of human inputs with at least p probability is derived through the MC and FIS modeling. Then a nonlinear control law is learned in the form of the neural architecture with CLBF loss function. The safety certificate is the existence of the CLBF loss value that satisfies the Lyapunov criteria (Fig. 7).

5.1 Automated Insulin Delivery Example

Data Description: We have accessed data from n = 20 patients with T1D for usage of the Tandem control IQ AID system for 22 weeks each (IRB information available). The patients were administered hydrocortisone dose of 40 mg, 20 mg, and 20 mg at 8 am, noon, and 2 pm on two supervised study days at study site.

Safety Violation in Control Affine Assumption: The illustrations in Fig. 8 is for an AID system, developed using the nonlinear optimal control theory discussed in Dawson *et al.* [11]. Data from FDA approved T1D simulator is used to train a neural network with the CLBF loss function (code available in `MIT-REALM/neural_clbf`). The multi-layer perceptron network was trained using 20,000 simulation data points to learn a CLBF. The set of initial glucose [110 mg/dl − 140 mg/dl] showed the $\mathbb{FI}$ property in absence of manual inputs, since the neural CLBF controller always keeps state trajectories within the initial set (shown by gray band in Fig. 8). CLBF safety certificate generation mechanisms (Step 1–4 in Fig. 8) [11] applied to AID system assumes the T1D patient as external with meal and correction bolus insulin as independent identically distributed (i.i.d) random disturbances. A safety certified AID has large glucose excursions due to meal intake. However, it will “eventually” enter a subset of the

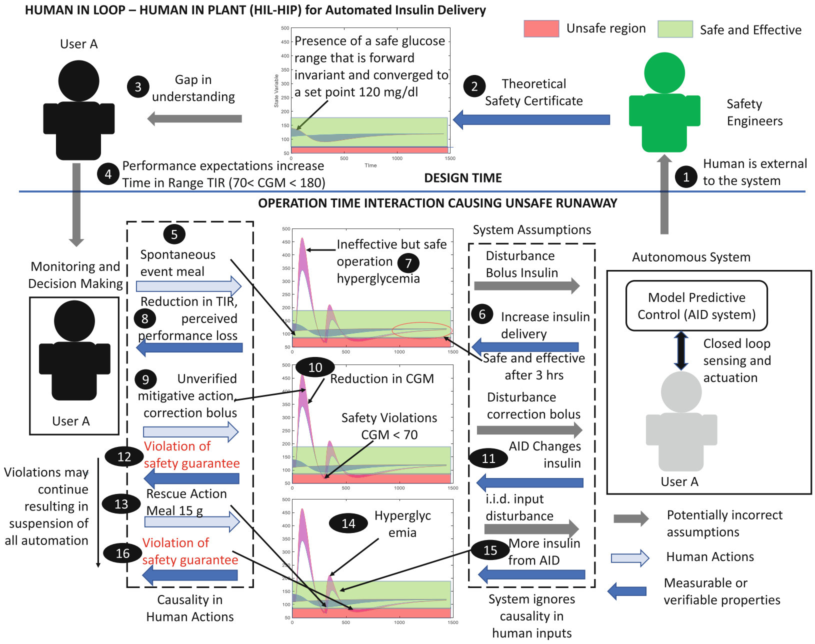

HUMAN IN LOOP – HUMAN IN PLANT (HIL-HIP) for Automated Insulin Delivery
Unsafe region
Safe and Effective
User A
3
Gap in understanding
Presence of a safe glucose range that is forward invariant and converged to a set point 120 mg/dl
2
Theoretical Safety Certificate
Safety Engineers
4
Performance expectations increase Time in Range TIR (70< CGM < 180)
DESIGN TIME
1
Human is external to the system
OPERATION TIME INTERACTION CAUSING UNSAFE RUNAWAY
Monitoring and Decision Making
User A
5
Spontaneous event meal
Reduction in TIR, perceived performance loss
8
9
Unverified mitigative action, correction bolus
12
Violation of safety guarantee
13
Rescue Action Meal 15 g
16
Violation of safety guarantee
Violations may continue resulting in suspension of all automation
Causality in Human Actions
Ineffective but safe operation
7
hyperglycemia
10
Reduction in CGM
Safety Violations CGM < 70
14
Hyperglyc emia
System Assumptions
Disturbance Bolus Insulin
6
Increase insulin delivery
Safe and effective after 3 hrs
Disturbance correction bolus
AID Changes insulin
11
i.i.d. input disturbance
15
More insulin from AID
System ignores causality in human inputs
Autonomous System
Model Predictive Control (AID system)
Closed loop sensing and actuation
User A
Potentially incorrect assumptions
Human Actions
Measurable or verifiable properties
State Variable
Time
500
450
400
350
300
250
200
150
100
50
0
500
1000
1500

Fig. 8. Safety violations occur if human inputs are considered as external disturbances in AID systems for T1D.

state space that has the 𝔽𝕀 property and hence be safe (no hypoglycemia glucose > 70 mg/dL). For a meal input of 100 g at 30 min with 2U of rapid acting insulin suggested by the bolus wizard [1] using ISF and CR settings, the system "eventually" (i.e. after 3 h) brings the glucose to 120 mg/dl (the setpoint), pink band in Fig. 8. According to the ADA recommendations and physicians advice, 2 h after meal CGM should be within safe range, however, the patient observed CGM to be above 300 mg/dl. The patient following clinician guidelines came up with a plan of using a correction bolus computed as: $CB = (300 - 120)/CF$, for a correction factor (CF) of 20 mg/U, resulting in $9U$ of insulin bolus. This resulted in the trajectory that cause hypoglycemia (steps 5 -7, intersection of pink band with red region in Fig. 8).

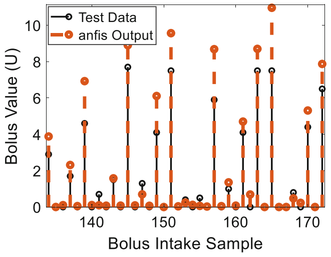
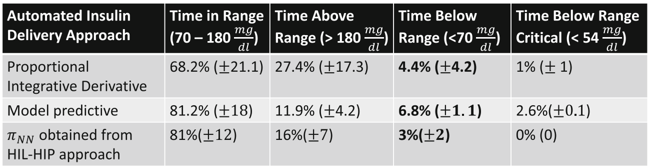


Fig. 9. ANFIS bolus prediction.

However, these interventions do not have safety guarantees and hence can lead the system to unsafe states (hypo-glycemia, glucose < 70 mg/dL in Steps 8–12 in Fig. 8). An unsafe excursion (CGM < 70 mg/dl) prompts the human to take immediate rescue carbohydrate of 15 g which drives up the glucose but takes it above 180 mg/dL, when the user could decide on getting another correction bolus. This can continue in operation time resulting in interaction runaway scenarios. The AS operation must be suspended to fail-safe modes. This is seen in nearly all AID systems such as Tandem Control IQ [8]. Utilizing the theory presented in the paper, we use the MPC Control IQ strategy as π_{nom} and learn π_{NN} to avoid hypoglycemia in presence of human inputs.

FIS Model Accuracy: We utilized an ANFIS [22] to predict individual external insulin bolus intake. There were an average of 522 (±15) meals and meal boluses and 261 (±30) correction bolus without meal. The ANFIS was designed with a 3×1 input vector consisting of {mean CGM, insulin on board computed using FIASP insulin action curve [8], carbohydrate intake}, and the output value of insulin bolus. With an 80-20 train-test split, the ANFIS achieved an RMSE of 0.45 in predicting insulin bolus (Fig. 9 shows prediction results). Execution time of the FIS model learning using the Matlab R2022 toolbox was 652 s on an Intel Core i7 processor.

MC Modeling Accuracy: The MC had three states {Large, Medium, Small} indicating meal sizes. The T1D dataset was used to obtain three clusters of meal sizes for each individual. Utilizing 522 meal instances, the transition probabilities of the 3 state MC was learned. Monte Carlo simulation of the MC gave the distribution of each meal size, which matched with the distribution in real data ($p = 0.041$). Execution time of MC learning was 0.267 s.



Automated Insulin Delivery Approach	Time in Range $(70 - 180\,\frac{mg}{dl})$	Time Above Range $(> 180\,\frac{mg}{dl})$	Time Below Range $(<70\,\frac{mg}{dl})$	Time Below Range Critical $(< 54\,\frac{mg}{dl})$
Proportional Integrative Derivative	68.2% (±21.1)	27.4% (±17.3)	**4.4% (±4.2)**	1% (± 1)
Model predictive	81.2% (±18)	11.9% (±4.2)	**6.8% (±1. 1)**	2.6%(±0.1)
π_{NN} obtained from HIL-HIP approach	81%(±12)	16%(±7)	**3%(±2)**	0% (0)

Fig. 10. Hypoglycemia reduction with HIL-HIP model.

Performance of Safety Certified HIL-HIP System. π_{nom} was developed as a Model Predictive Control (MPC). A Bayesian meal prediction scheme utilizing the MC was integrated with the FIS to obtain the meal and correction bolus information. The integrated HIL-HIP system was simulated in closed loop with a Python implementation of Runge-Kutta solution of the BMM. The Neural CLBF architecture was then used with the MPC + Bayesian + FIS control as π_{nom} which learned a CLBF and a new controller π_{NN}. The neural CLBF architecture is an MLP with 128 hidden layers. The input dimension is 3×288, where a single day CGM, Interstitial insulin and blood insulin was delivered as input. The sigmoid activation function was used in each neuron. The neural CLBF MLP was trained for 200 epochs which took 22 h on a 8 core Intel i7 CPU. We compare performance of π_{NN} with regular MPC and Proportional Integrative and Derivative (PID) using 5 subjects in the T1D simulator [19] in Fig. 10, which shows significant reduction in hypoglycemia for π_{NN}.

5.2 Autonomous Braking Example

Dataset Description: We evaluated the ABS example in simulation. We considered a driver who has two reaction time distributions: a) low reaction, mean time 300 ms, with SD 25 ms, and b) high reaction time, mean time 500 ms, with SD 125 ms. Each type of reaction time is sampled from a Gaussian distribution.

The braking behavior results in two types of manual braking force: a) hard braking, with a step braking force of $u_h = 5\,\mathrm{m/s}^2$ until the car stops, and b) soft braking, with a step braking force of $u_h = 3\,\mathrm{m/s}^2$ until the car stops.

Simulation Setup to Generate Data: We developed a two car Simulink model that simulates Eq. 3 for various initial distance and initial relative velocities (Fig. 11). The Simulink model consists of two parts: a) Braking control, and b) Vehicle kinematics simulator. Two types of braking control is simulated: i) linear quadratic Gaussian (LQG) as baseline comparator, and ii) our proposed neural CLBF based controller that takes into account human inputs. The human action model is a combination of an MC model for the reaction time and a FIS model for the braking behavior, both of which are implemented as Matlab functions.

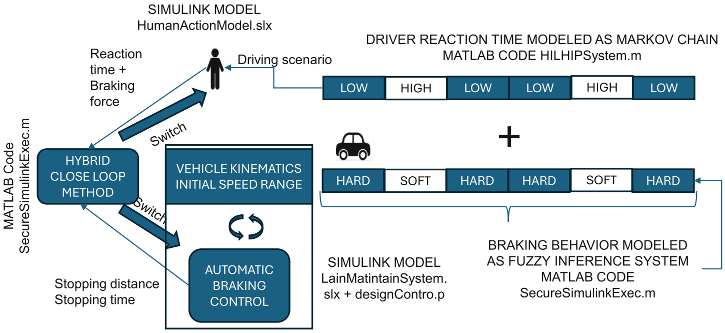


Fig. 11. Simulation setup for the ABS example.

Data Generation: We generated 200 driving scenarios. The initial distance and velocity were derived from a Gaussian distribution. The MC model of human reaction time (detailed below) was simulated using the Monte Carlo method [25] for 200 samples which provided a sequence of 200 low and high reaction times. For each test case, the FIS model (detailed below) was simulated for 200 samples to obtain a braking force decision. One test case consisted of one initial distance, initial velocity, human reaction time and human braking force. The hybrid close loop operation of the autonomous braking system was implemented using a matlab code that decided on the human intervention based on the distance between the cars and reaction time of the user. To evaluate the learned controller and baselines, 200 test cases were generated that were never used to learn the FIS or MC models.

FIS Modeling: An ANFIS model is developed to predict the hard and soft braking results. We utilized 120 samples to train the ANFIS model and used 80 samples to test. The RMSE in estimating the braking force was 0.01 m/s^2.

MC Modeling: The transition between the two reaction times is modeled using a Markov chain with two states. The MC model was trained using 60 samples from the 200 generated data. After 60 samples, the MC model reached steady state and hence was stopped. The accuracy of the MC model in determining low or high reaction time was 87% in the rest of 140 samples.

Evaluation Metrics: The baseline LQG controller and the neural CLBF learned controller are evaluated in terms of percentage number of avoided collisions out of the 200 test cases.

Performance of Safety Certified HIL-HIP ABS Controller: π_{nom} was developed as a LQG controller. The integrated HIL-HIP system was simulated in closed loop with a Python implementation of Runge-Kutta solution of the kinematics model in Eq. 3. The Neural CLBF architecture was then used with the LQG + MC + FIS control as π_{nom} which learned a CLBF and a new controller π_{NN}^{ABS}. The neural CLBF architecture is an MLP with 128 hidden layers. The input dimension is 3×288, where a relative distance, relative velocity and car A acceleration was delivered as input. The sigmoid activation function was used in each neuron. Similar training regimen as the AID example was used for ABS. We compare performance of π_{NN}^{ABS} with regular LQG controller as the comparator.

The learned controller π_{NN}^{ABS} avoided 99% of collisions out of the 200 test samples. While the LQG controller only had a collision avoidance rate of 84%.

6 Related Works

Three broad classes of controller synthesis exist- **a) Optimization approach:** For linear systems with eventual guarantees, a LQG optimal control strategy exists [14], which guarantees that a safety related STL will be satisfied. For nonlinear systems with eventual guarantees, control Lyapunov (CLF) theory exists [21], which guarantees safety in absense of human inputs.

b) Game theoretic approach: The controller synthesis problem has been modeled as a two player game between the environment and the controller for safe HIL control [16]. These methods work well for 1D decision problems such as detection of safe switching time.

c) Reinforcement learning approach: Safe RL is an emerging approach that models agents with a value function that has control objective as the reward and safety violation as the penalty function [13]. Safe RL technique starts an initial safe MPC design that may not be effective, and for each control step evaluates the value function. If the value function is less than a threshold indicating heavy penalty, the safe RL defaults to the MPC strategy, else it uses the strategy obtained by maximizing the value function.This approach has been frequently used in robotics, however, the value function evaluation strategy does not involve human inputs. To the best of our knowledge all attempts consider human inputs as external disturbances and as such may result in interaction runaway (Fig. 8).

7 Conclusions

The paper presents extensions of state-of-art safe controller synthesis theory that assumes humans as outside the system to enable controller synthesis for hybrid close loop systems with human modeled as a part of the system. Departing from the traditional approach of assuming human inputs as external disturbances, this paper considers the human and the autonomous system as a co-operating unified system. The novel integration of Markov chains with fuzzy inference system to model human control and the neural control architecture to synthesize safe controller provides a mechanism for developing HIL-HIP HCL based AS as a unified system. This can provide early feedback on the safety of the AS operation so that mitigative actions can be taken proactively to avoid fatal accidents. We show the application of the new theory on AID controller synthesis for T1D, where HIL-HIP AID is shows to outperform MPC and PID control with respect to safety and efficacy metrics and on Autonomous braking systems, where the HIL-HIP controller is much safer than the LQG control.

References

1. Andersen, A., et al.: Optimum bolus wizard settings in insulin pumps in children with type 1 diabetes. Diabet. Med. **33**(10), 1360–1365 (2016)
2. Banerjee, A., Gupta, S.K.S.: Your mobility can be injurious to your health: analyzing pervasive health monitoring systems under dynamic context changes. In: 2012 IEEE International Conference on Pervasive Computing and Communications (PerCom), pp. 39 –47 (2012). https://doi.org/10.1109/percom.2012.6199847
3. Banerjee, A., Gupta, S.K.: Analysis of smart mobile applications for healthcare under dynamic context changes. IEEE Trans. Mob. Comput. **14**(5), 904–919 (2015)
4. Banerjee, A., Maity, A., Kamboj, P., Gupta, S.K.S.: CPS-LLM: large language model based safe usage plan generator for human-in-the-loop human-in-the-plant cyber-physical system (2024). https://arxiv.org/abs/2405.11458. aI Planning for Cyber-Physical Systems - CAIPI'24 AAAI
5. Banerjee, A., Maity, A., Lamrani, I., Sandeep Gupta, K.: Co-operative game for certification and continued conformance check of AI enabled cps*. In: 2024 IEEE 7th International Conference on Industrial Cyber-Physical Systems (ICPS), pp. 1–6 (2024). https://doi.org/10.1109/ICPS59941.2024.10639951
6. Banerjee, A., Venkatasubramanian, K.K., Mukherjee, T., Gupta, S.K.S.: Ensuring safety, security, and sustainability of mission-critical cyber-physical systems. Proc. IEEE **100**(1), 283–299 (2011)
7. Banerjee, A., Zhang, Y., Jones, P., Gupta, S.: Using formal methods to improve home-use medical device safety. Biomed. Instrum. Technol. **47**(s1), 43–48 (2013)
8. Breton, M.D., Kovatchev, B.P.: One year real-world use of the control-IQ advanced hybrid closed-loop technology. Diab. Technol. Ther. **23**(9), 601–608 (2021)

9. Chen, X., Ábrahám, E., Sankaranarayanan, S.: Flow*: an analyzer for non-linear hybrid systems. In: Computer Aided Verification: 25th International Conference, CAV 2013, Saint Petersburg, Russia, 13–19 July 2013. Proceedings 25, pp. 258–263. Springer (2013)
10. Cohen, M.B., Lee, Y.T., Song, Z.: Solving linear programs in the current matrix multiplication time. J. ACM **68**(1) (2021). https://doi.org/10.1145/3424305
11. Dawson, C., Gao, S., Fan, C.: Safe control with learned certificates: a survey of neural lyapunov, barrier, and contraction methods. arXiv preprint arXiv:2202.11762 (2022)
12. Donzé, A., Maler, O.: Robust satisfaction of temporal logic over real-valued signals. In: Formal Modeling and Analysis of Timed Systems: 8th International Conference, FORMATS 2010, Klosterneuburg, Austria, 8–10 September 2010. Proceedings 8, pp. 92–106. Springer (2010)
13. Garcıa, J., Fernández, F.: A comprehensive survey on safe reinforcement learning. J. Mach. Learn. Res. **16**(1), 1437–1480 (2015)
14. Karaman, S., Sanfelice, R.G., Frazzoli, E.: Optimal control of mixed logical dynamical systems with linear temporal logic specifications. In: 2008 47th IEEE Conference on Decision and Control, pp. 2117–2122. IEEE (2008)
15. Lamrani, I., Banerjee, A., Gupta, S.K.S.: Certification game for the safety analysis of AI-based CPS. In: Habli, I., Sujan, M., Gerasimou, S., Schoitsch, E., Bitsch, F. (eds.) Computer Safety, Reliability, and Security. SAFECOMP 2021 Workshops, pp. 297–310. Springer, Cham (2021)
16. Li, W., Sadigh, D., Sastry, S.S., Seshia, S.A.: Synthesis for human-in-the-loop control systems. In: Tools and Algorithms for the Construction and Analysis of Systems: 20th International Conference, TACAS 2014, Held as Part of the European Joint Conferences on Theory and Practice of Software, ETAPS 2014, Grenoble, France, 5–13 April 2014. Proceedings 20, pp. 470–484. Springer, Cham (2014)
17. Maity, A., Banerjee, A., Gupta, S.: Detection of unknown-unknowns in cyber-physical systems using statistical conformance with physics guided process models. arXiv preprint arXiv:2309.02603 (2023)
18. Maity, A., Banerjee, A., Gupta, S.K.: Detection of unknown-unknowns in human-in-loop human-in-plant systems using physics guided process models. In: 2023 57th Asilomar Conference on Signals, Systems, and Computers, pp. 1500–1504 (2023). https://doi.org/10.1109/IEEECONF59524.2023.10476736
19. Man, C.D., Micheletto, F., Lv, D., Breton, M., Kovatchev, B., Cobelli, C.: The UVA/PADOVA type 1 diabetes simulator: new features. J. Diabetes Sci. Technol. **8**(1), 26–34 (2014)
20. Priyanka Bagade, Ayan Banerjee, S.K.G.: Validation, verification, and formal methods for cyber-physical systems. In: Cyber-Physical Systems, pp. 175–191. Elsevier (2017)
21. Richards, S.M., Berkenkamp, F., Krause, A.: The lyapunov neural network: adaptive stability certification for safe learning of dynamical systems. In: Conference on Robot Learning, pp. 466–476. PMLR (2018)
22. Salleh, M.N.M., Talpur, N., Hussain, K.: Adaptive neuro-fuzzy inference system: overview, strengths, limitations, and solutions. In: Data Mining and Big Data: Second International Conference, DMBD 2017, Fukuoka, Japan, 27 July–1 August 2017, Proceedings 2, pp. 527–535. Springer, Cham (2017)
23. Seshia, S.A., Sadigh, D., Sastry, S.S.: Towards verified artificial intelligence. arXiv preprint arXiv:1606.08514 (2016)

24. Welch, S., Gebhart, S., Bergman, R., Phillips, L.: Minimal model analysis of intravenous glucose tolerance test-derived insulin sensitivity in diabetic subjects. J. Clin. Endocrinol. Metab. **71**(6), 1508–1518 (1990)
25. Ávila, D., Junca, M.: On reachability of Markov chains: a long-run average approach. IEEE Trans. Autom. Control **67**(4), 1996–2003 (2022). https://doi.org/10.1109/TAC.2021.3071334

Multi-frequency Fine-Grained Matching for Audio-Visual Segmentation

Yinhao Zhang[1], Tianyang Xu[1], Xiao-Jun Wu[1](✉), Shaochuan Zhao[1], and Josef Kittler[2]

[1] School of Artificial Intelligence and Computer Science, Jiangnan University, Wuxi, People's Republic of China
{yinhaozhang,7201905026}@stu.jiangnan.edu.cn, {tianyang.xu,wu_xiaojun}@jiangnan.edu.cn

[2] Centre for Vision, Speech and Signal Processing, University of Surrey, Guildford GU2 7XH, UK
j.kittler@surrey.ac.uk

Abstract. Audio-visual segmentation is a recently proposed task, whose main goal is to locate the target of the sound in the image at the pixel level. In practical scenarios, multiple types of audio can coexist, with different objects emitting sounds at different frequencies. However, existing methods only use single-frequency audio information when fusing audio and visual modalities. Moreover, the process of combining images and audio can be quite rough. Therefore, we propose a multi-frequency fine-grained matching method for multiple sound sources scenario. Firstly, we use short-time Fourier transform (STFT) to extract different frequency spectrograms and input them into the audio encoder to extract multi-frequency audio features. Secondly, multi-frequency audio information serves as a prompt in the pixel decoder stage to guide model segmentation. To obtain high-quality prompts, we use an attention method in the Audio-Visual Matching Module (AVMM) to match visual and audio information. The experiments show that our method has a significant improvement over the baseline and achieves state-of-the-art results on the MS3 benchmark (64.1 mIoU on MS3).

Keywords: Audio-Visual Segmentation · Multi-Modality · Prompter

1 Introduction

Visual and audio information plays an important role in helping human beings understand the real world. In natural scenarios, multiple modalities of information exist, audio modality can provide important data that cannot be captured visually. Therefore, based on single-modality video segmentation, audio information is added to form a new task called: Audio-Visual Segmentation(AVS) [35]. The goal of this task is to segment the sound-emitting target in video frames at the pixel level. This task is divided into two subsets: Single Sound Source Segmentation (S4) and Multiple Sound Source Segmentation (MS3).

A. Antonacopoulos et al. (Eds.): ICPR 2024, LNCS 15329, pp. 36–50, 2025.
https://doi.org/10.1007/978-3-031-78110-0_3

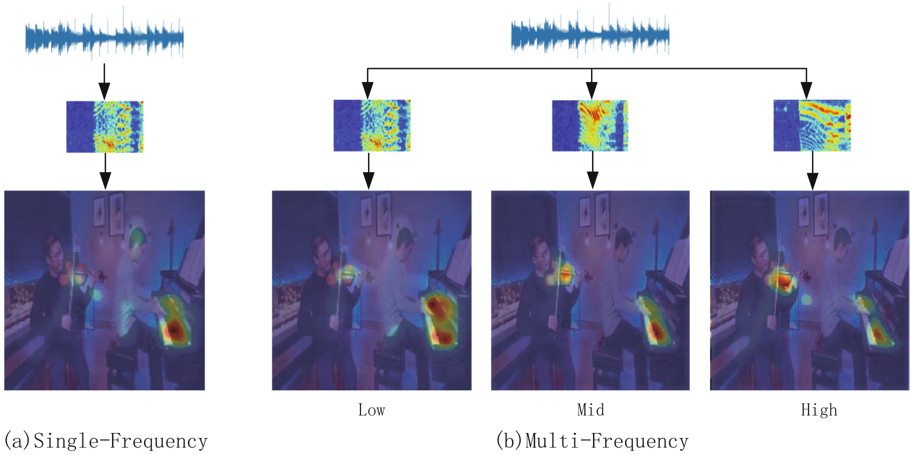


Fig. 1. Compared with previous methods, we use multi-frequency audio information to improve the model's recognition of different targets. (a) Previous methods only used a single type of frequency audio for the image and audio modality matching process. Different audio frequencies are produced by different sound source targets. Therefore, using single-frequency audio for differentiation recognition of different targets is relatively weak (b) Our method employs multi-frequency audio features to discern different targets. As the audio frequency changes, the model's attention changes accordingly.

For the task of S4, Zhou *et al.* [35] proposed to use the cross-attention method to fuse audio signals and image clues, which achieved good segmentation results. Subsequently, a series of audio-visual works were proposed [9,12,17,18,24], which further improved segmentation performance. In these works, only single-frequency audio information is utilised. However, in a real-world scene, the audio emitted by targets are different in decibels, timbre, and frequency. In MS3 task, there exists multiple sound source targets simultaneously. Therefore, the above methods for MS3 tasks are not entirely applicable for S4 tasks. Therefore, we employ multi-frequency audio features to improve the model's recognition of different targets. In addition, existing methods are not meticulous enough in the matching and fusion process of the audio and image modalities, causing the model to be more inclined to recognize salient objects rather than sound source objects. Therefore, how to use audio information to accurately guide the model's understanding of sound source objects is a significant challenge.

To tackle the above mentioned problems, we first try to analyse how different frequency audios affect the way the model focuses on targets. Taking the left side of Fig. 1 as an example, for single-frequency audio feature, model's attention towards the violin is not enough. Whereas on the right side of Fig. 1, with the increase in frequency, the model's attention shifts from the piano to the violin. Therefore, in a multi-frequency audio mixing scenario, we increase the granularity of the frequency to improve the discrimination of individual targets. To be specific, we choose audio features of different frequencies as the input to the model, which provides sufficient clues for the matching of the target appearance and its corresponding sound source.

Moreover, in order to promote the fine-grained fusion of image and audio features. We initialize audio prompt through AVMM which adaptively aligns the image clues and audio features. For we have extracted multi-frequency audio features, the AVMM allows the fine-grained matching of the audio signal and the target appearance. We take the audio prompt to guide the model's learning of the sound source target in the pixel decoder stage. In this way, the problem of mixing multiple audio signals in complex scenarios is solved. Our proposed method is more effective for MS3 tasks. To summarise, the main contributions of our work are summarized as follows:

- We propose a simple but effective strategy where we change the frame length in the STFT to obtain audio features of different frequencies, aiming to distinguish different sound sources in a mixed audio scenario.
- We narrow the gap between the alignment of the audio and visual modalities through AVMM and audio prompt module. We introduce rich audio features and continuously change prompts at different stages of the pixel decoder.
- The extensive experiments on the AVS task verify that our method can efficiently utilize multi-frequency audio information, demonstrating that our method significantly outperforms existing baseline methods. At the same time, it achieves new state-of-the-art performance on the AVS task, *e.g.*, 64.1% mIoU with PVT backbone on AVSBench(MS3) benchmark.

2 Related Works

2.1 Semantic Segmentation

Semantic segmentation is a pixel-level classification task within an image region. Semantic segmentation can distinguish the foreground and background in an image and classify and identify each object. Currently, the research on semantic segmentation methods is generally divided into two types: CNN-based [2,22,33] and transformer-based [4,28,31,34] segmentation methods. In recent years, with the advancement of Convolutional Neural Network (CNN), the Fully Convolutional Networks (FCN) proposed by Long *et al.* [22] uses convolutional layers to replace the original fully connected layers for pixel level end-to-end prediction. In order to solve the problem of the information loss brought by the decrease of the spatial resolution, Chen *et al.* [2] designed DeepLab, which significantly improved FCN by incorporating atrous convolution in the decoder and utilizing bilinear interpolation for upsampling. After adopting the visual transformer(VIT) [5] method, global feature information is captured through its unique attention mechanism. However, in dense prediction tasks, the computational cost is significantly increased. Subsequently, a series of works followed this approach. Among them, Swin Transformer [21] uses a moving window method to reduce the amount of computation, but it is still very time-consuming. SegFormer [31] removes the position encoding and complex decoder, which alleviates high computational cost of VIT [5] and SETR [34] for large images. These segmentation methods have made significant contributions to solving the AVS task.

2.2 Audio-Visual Segmentation

With the development of multi-modality, lots of audio-visual tasks have been proposed. Such as, sound source localization(SSL) [1,26,27], audio-visual segmentation(AVS) [9,12,15,17,18,24,35] and audio-visual spatialization [6,25]. In this paper, we focus on the task of audio-visual segmentation which proposes to segment target objects by their sound. To enhance the audio-visual segmentation accuracy, recent methods [3,7,15,23,32,35] use transformer-based cross-modal fusion strategy. Zhou *et al.* [35] built an audio-visual segmentation benchmark and proposed a Temporal Pixel-wise Audio-Visual Interaction module(TPAVI). This method encodes the spatio-temporal audio-visual interaction of the entire video at the pixel level, narrowing the gap between the image modality and the audio modality. However, the design of the global modality fusion strategy is too rough for the matching of audio and visual information. We consider that the problem of fine-grained alignment between audio and visual modalities is a crucial challenge to AVS task. To this end, Li *et al.* [15] designed Decoupled Audio-Visual Transformer Encoding Module (DAVT), which uses the pixel-level fusion module Blockwise-Encoded Gate(BEG) to fuse the audio and visual features of the corresponding frames. However, when there are multiple sound source objects in the same frame, this method does not distinguish sufficiently between different targets, leading to suboptimal segmentation results. Therefore, distinguishing different targets in mixed audio scenes is of great significance for the AVS task. So, AVSegFormer [7] is proposed to selectively focus on the visual features of interest by directly introducing audio features into the encoder-decoder architecture. However, when multiple sound source targets in the same frame, the discrimination the between targets is weak. Based on this observation, we focus on the differences in sound frequencies produced by different targets. We proposed a multi-frequency fine-grained method based on [7] considering the richness of audio information itself for AVS tasks, and introduced AVMM and pixel decoder to guide the model to recognize the sound source target in a more detailed manner based on audio information.

2.3 Visual Prompt Learning

The fine-tuning strategy is to pre-train the model without significantly changing the model structure and parameters. By adding prompt information to the model input, the model itself can solve the problem. Therefore, only a few parameters are needed to align the pre-trained model to downstream tasks. Kirillov *et al.* [13] designed the benchmark Segment Anything (SAM), which use point, box, mask, and text to form prompts through prompt encoder, and achieves good results in downstream tasks such as zero-shot learning and edge detection.

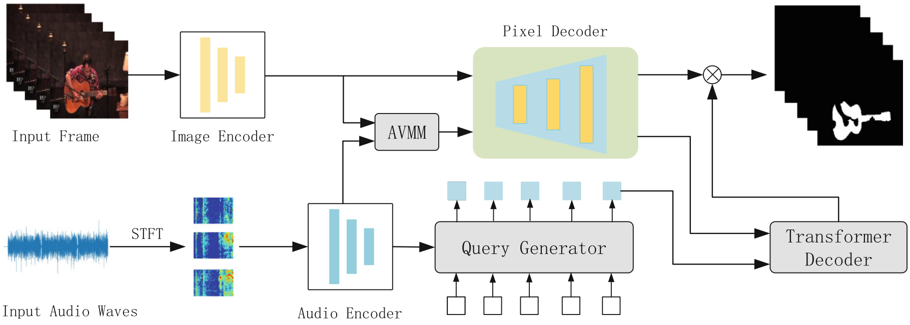


Fig. 2. Illustrations of the overall architecture. The image encoder and audio encoder extract multi-scale feature maps and audio features respectively. AVMM matches audio features with fine-grained images. The pixel decoder guides image audio-visual recognition via efficient audio information. The query generator initializes queries with audio information. Finally, the transformer decoder aims to decouple the sound source objects in the image.

Li *et al.* [14] designed the Semantic-SAM model, which generates semantically-aware multi-granularity masks using a query-based mask decoder, optimizing the quality of the output masks. Wang *et al.* [30] constructs semantically-aware audio prompts to bridge the semantic gap between visual and auditory modalities. It uses relevance adapters to preserve the prior knowledge of the visual base model and explores the generalization of AVS tasks through prompts. In addition, the prompt learning method can also promote multi-modal fusion. Zhu *et al.* [36] proposed VIPT, a multi-modal tracking learning framework using visual prompt-tuning. Based on this idea, we designed modality-complementary prompter (MCP) and pixel decoder to explore the audio-visual segmentation task through more detailed audio-visual matching.

3 Methodology

3.1 Overview

In this section, we first describe the overall structure of the model and then present the details of the component modules. Overall, the audio-visual fine-grained matching method has three sub-modules, *i.e.*, multi-frequency audio extraction, AVMM and pixel decoder. The overall structure of our proposed method is illustrated in Fig. 2. Multi-frequency audio features are obtained through multi-frequency audio extraction, and together with visual features, they are input into AVMM for fine-grained matching. In the Pixel decoder, audio features are used to guide visual information to focus on the sound source target by prompting. Following the existing segmentation methods [28,31,33], we chose the encoder-decoder structure and selected AVSegFormer [7] as the baseline of our method. We chose the VGGish [11] which was pre-trained on

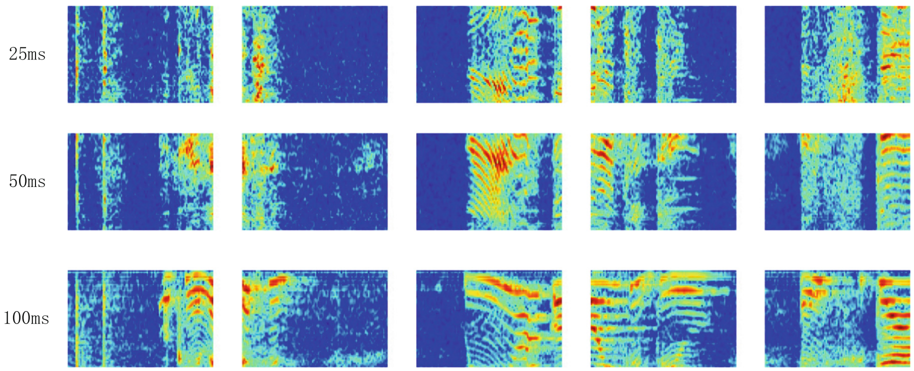


Fig. 3. The figure shows the visualization of different frequencies in the audio file. To generate audio frequency differences for the sound source targets. We converting the audio signal into a spectrogram through STFT, frame lengths of 25 ms, 50 ms, and 100 ms are selected respectively. The frequency of the spectrogram increases simultaneously with the selected frame length.

AudioSet [8] dataset as the audio encoder to extract audio features. ResNet-50 [10] and PVTv2 [29] were chosen as image encoder to extract multi-scale visual features.

3.2 Multi-frequency Audio Feature

We follow the VGGish [11] method to extract audio features. Initially, we resample the audio signal to 16KHz mono, and then separate the speech frames according the window length and the periodic Hann window. In previous methods, spectrogram $\hat{A}_{mel} \in \mathbb{R}^{T\times 96\times 64}$(64 mel-spaced frequency bins over 96 time steps) was obtained using STFT with a window of 25ms and a hop size of 10ms. In order to extract multi-frequency spectrogram $A_{mel} \in \mathbb{R}^{N_{fre}\times T\times 96\times 64}$, we select N_{fre} kinds of frame length. Finally, the mel spectrograms are input into VGGish [11] to obtain the audio feature $F_a \in \mathbb{R}^{T\times N_{fre}\times D}$. T and D represents the number of frames and the audio feature dimension. In this paper, we selects frame length of 50ms and 100ms as additional multi-frequency audio signals. The visualization results of the mel spectrograms of different frequencies are shown in Fig. 3. We expand the frequency as a new dimension to provide more sufficient audio information.

3.3 Audio-Visual Matching Module

The main purpose of this module is to enable fine-grained adaptive matching of multi-frequency audio features and visual features. We obtain the multi-scale feature map $\mathcal{F}_{visual}$ through the visual encoder. $\mathcal{F}_{visual}$ can be written as $\mathcal{F}_{visual} = \{\mathcal{F}_1, \mathcal{F}_2, \mathcal{F}_3, \mathcal{F}_4\}$, where $\mathcal{F}_i \in \mathbb{R}^{C_i\times \frac{H}{2^{i+1}}\times \frac{W}{2^{i+1}}}$ and $i \in [1, 2, 3, 4]$. The last three feature maps $\widehat{\mathcal{F}}_{visual}$ are selected as subsequent inputs. $\widehat{\mathcal{F}}_{visual}$

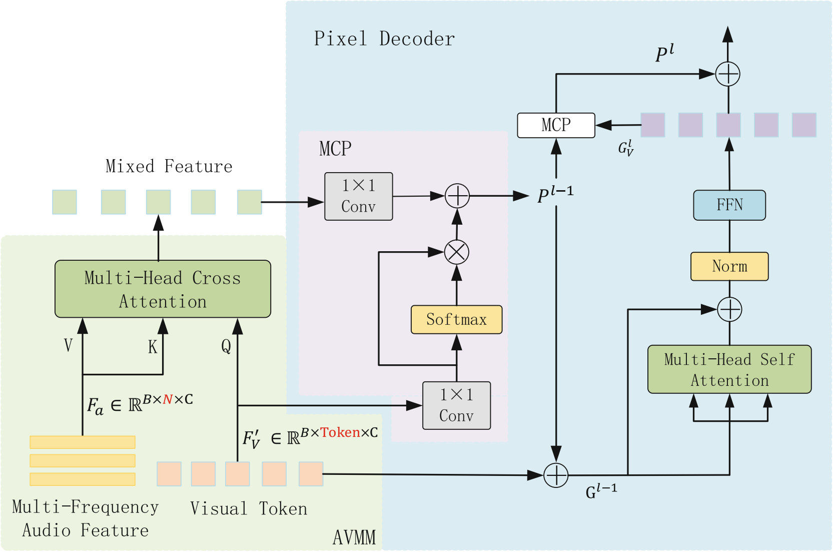

Pixel Decoder
P^l
MCP
G_V^l
MCP
1×1 Conv
P^{l-1}
Mixed Feature
FFN
Norm
Softmax
Multi-Head Cross Attention
V
K
Q
Multi-Head Self Attention
1×1 Conv
$F_a \in \mathbb{R}^{B\times N\times C}$
$F_V' \in \mathbb{R}^{B\times Token\times C}$
G^{l-1}
Multi-Frequency Audio Feature
Visual Token
AVMM

Fig. 4. The architecture of the proposed multi-frequency fine-grained audio-visual segmentation method. This figure mainly shows the details of AVMM, MCP and pixel decoder. The audio and visual features are matched at a fine-granularity through AVMM to obtain mixed feature with the same dimension as the visual features. The MCP module combines the mixed feature and visual token to generate an audio prompt, and the pixel decoder is added to the AVS task to carefully guide the model to focus on the sound source target.

can be written as $\widehat{\mathcal{F}}_{visual} = \{\mathcal{F}_2, \mathcal{F}_3, \mathcal{F}_4\}$. Audio features F_a are extracted by VGGish after frequency channel expansion. The detail of AVMM is shown in Fig. 4. We flatten the feature maps $\mathcal{F}_V$ and concatenate to obtain visual token $F_V^{'} \in \mathbb{R}^{B\times token\times C}$. B and *token* represent batch size and visual token length respectively. We use $F_V^{'}$ as the query and F_a as the key and value to input into AVMM to obtain the mixed features F_M, which is computed as follow:

$$F_M = \text{AVMM}(F_v, F_a) = \text{softmax}(\frac{F_V W^Q (F_a W^K)^\mathsf{T}}{\sqrt{d_{head}}}) F_a W^V \tag{1}$$

$W^Q, W^K, W^V \in \mathbb{R}^{C\times d_{head}}$ represent the learnable parameters. The F_M are aligned with the visual token in dimension, further narrowing the gap between different modalities. This allows the visual token to select audio features that are more suitable for itself, increasing the differences between different sound source targets.

3.4 Pixel Decoder

Fine-tuning strategies have been widely used in many visual tasks [16,19,20]. Indeed, as the capacity and parameters of the deep model continue to increase, traditional fine-tuning methods struggle to balance the knowledge of the pre-trained model and the adaptation of the downstream task. Recently, prompt learning has emerged by fine-tuning a small number of parameters, and it has achieved excellent results, gradually becoming a new paradigm for fine-tuning. We believe that prompt learning can be used to narrow the gap between audio and visual modality. The module we introduces audio cues between each pixel decoder layers. The specific module is shown in Fig. 4. This method enhances the influence of audio made by the sound source target in the image. In order to enhance the tightness of different modal fusion, we refer to the MCP. As shown in Fig. 4, visual tokens and mixed features are input into the MCP to generate learned prompts $\mathcal{P}^{l-1}$. $\mathcal{P}^{l-1}$ and visual tokens are added to generate prompted tokens $\mathcal{G}^{l-1}$ and input into the pixel decoder layer. $\mathcal{G}_V^l$ is the output of the l-th pixel decoder layer. It is added to $\mathcal{P}^{l-1}$ through MCP to get $\mathcal{P}^l$. Finally, $\mathcal{G}_V^l$ and $\mathcal{P}^l$ are added element by element to get $\mathcal{G}^l$, which is computed as,

$$\mathcal{G}^l = \mathcal{G}_V^l + \mathrm{MCP}(\mathcal{P}^{l-1}, \mathcal{G}_V^l) \qquad l = 1, ..., L \tag{2}$$

where L represents the number of layers of pixel decoder and MCP.

4 Experiments

Through conducting experiments on AVSbench [35], we evaluated various audio-visual fine-grained matching methods for the segmentation of sound source from multiple targets. We also conducted ablation experiments on the proposed modules to evaluate the impact of proposed network modules in our study on sound source target detection.

4.1 Experimental Setup

Datasets. MS3 and S4 are the two main subsets of the dataset. In the S4 subset, each video has only one sound source target. There are 4932 videos in total, and the number of training, validation, and test sets is 3452/740/740 respectively. MS3 refers to a dataset in which multiple sound source targets appear sequentially or simultaneously in a video. The MS3 dataset has a total of 424 videos, including 296 training sets, 64 validation and test sets. In AVSbench benchmark, each video lasts for five seconds, and only the first frame of every second is taken. During training, S4 only has annotated for the first frame, unlike S4, MS3 has annotated for five frames during training.

Implementation Details. Following [3,7,32], we use ResNet-50 and PVTv2 pre-trained on the ImageNet as the backbone for extracting image features. Image frames are resized to 224×224. The number of decoder layers and queries settings follow the baseline. We apply AdamW optimizer with a learning rate of 2×10^{-5} and set batch size of 2. The MS3 dataset needs to be trained for 60 epochs. 30 epochs are trained on the S4 dataset. Only a single NVIDIA 3090 GPU is needed to train our model.

Metrics. Following previous works [7,15,32], we selected Jaccard index $\mathcal{J}$ and F-score $\mathcal{F}$ as the evaluation indicators of experimental results. $\mathcal{M_J}$ represents mean intersection over union (mIoU), and $\mathcal{M_F}$ represents mean precision and recall: $F_\beta = \frac{(1+\beta^2)\times precision \times recall}{\beta^2 \times precision + recall}$, where β^2 set to 0.3 in our experiments.

4.2 Main Results

Table 1. Comparison with sota methods on the MS3 subset of AVSbench

Methods	ResNet-50		PVT-v2	
	$\mathcal{M_J}$	$\mathcal{M_F}$	$\mathcal{M_J}$	$\mathcal{M_F}$
AVSBench [35]	47.9	57.8	54.0	64.8
AVS-BiGen [9]	45.0	56.8	55.1	66.8
BAVS [3]	50.2	62.4	58.6	65.5
COMBO [32]	54.5	66.6	59.2	71.2
ECMVCE [24]	48.7	60.7	57.8	70.8
DiffusionAVS [23]	49.8	58.2	58.2	70.9
AVSegFormer [7]	49.53	62.8	58.4	69.3
Ours	**56.7**	**70.2**	**64.1**	**73.9**

Performance Comparison. AVS [35] was proposed as an innovative task, with the purpose of segmenting objects that make sounds in video frames. We have collected the most recent methods for AVS task, the results of which are presented in Table 1. These methods include AVSBench [35], AVS-BiGen [9], BAVS [3], COMBO [32], ECMVCE [24], DiffusionAVS [23], and AVSegFormer [7]. Table 1 presents a comparison of the experiments on the MS3 subset, showing an improvement of 7.17$\mathcal{M_J}$ and 7.4$\mathcal{M_F}$ compared to baseline method on ResNet-50 backbone. In the instance of using PVTv2 as the backbone, our method showed an improvement of 4.1$\mathcal{M_J}$ and 2.3$\mathcal{M_F}$ compared to the SOTA methods. Qualitative experiments have shown that our method can effectively deal with scenarios where multiple audio signals are mixed, distinguishing between different sound sources.

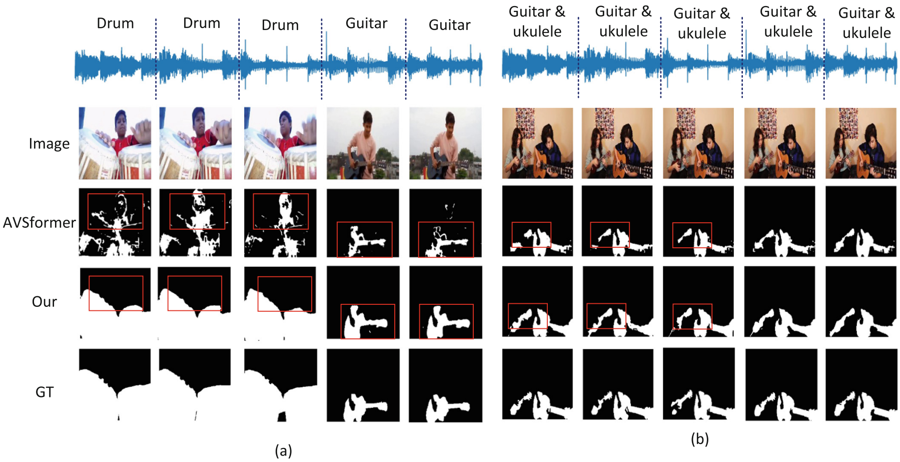


Fig. 5. Qualitative visualization of the audio-visual segmentation task. The proposed multi-frequency fine-grained matching produces more accurate and higher quality mask.

Masks Visualization. Figure 5 elucidates the visualization results of our experimental results on MS3 subsets, contrasting with the baseline methods. The MS3 subset contains multiple sound source targets that can change over time. Indeed, as shown in Fig. 5(a), the sound source target transitions from drums to guitar. Our method can accurately segment sound source targets by considering the variations in audio. On the other hand, AVSegFormer [7] fails to correctly identify the source target during the segmentation process. This demonstrates that our method is based on audio information to locate targets, rather than finding conspicuous targets in the image. For MS3 task, as depicted in Fig. 5(b), both AVSegFormer and our proposed method can localize the sound source target. However, our method offers higher quality detail description and mirrors the ground truth more closely. There are sounds from two different musical instruments simultaneously. AVSegFormer cannot capture the details of the ukulele in the first three frames, indicating difficulty in distinguishing target differences in mixed audio scenarios. In such mixed audio scenarios, guidance from audio information is increasingly required. It's evident that we can fully segment out different sound source targets and maintain full detail, leading to high-quality masks.

4.3 Ablation Study

In this section, we verify the effectiveness of each component of the proposed audio-visual fine-grain matching method. In the ablation experiments, we used the PVTv2 [29] as the visual backbone on the MS3 subset. Firstly, we analyzed the impact of audio prompter on the task of audio-visual segmentation. Comparatively, the information from audio prompter guiding the model's seg-

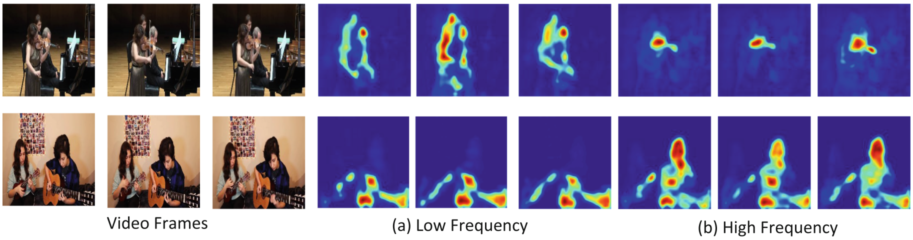


Fig. 6. Visualize the attention map to explore the relationship between the sound source target and audio frequency. The left side shows the visual frames. (a) and (b) represent using low-frequency audio and high-frequency audio as inputs to the model to find the corresponding sounding objects respectively.

mentation significantly improved the baseline, with $\mathcal{M}_{\mathcal{J}}$ (mIoU) rising from the original 54.0 to 57.1, and $\mathcal{M}_{\mathcal{F}}$ from 64.5 to 67.7, with results as shown in Table 2. Secondly, we further investigated the impact of the audio-visual matching strategy. By incorporating this module, $\mathcal{M}_{\mathcal{J}}$ increased by 3.3. This means that more detailed calibration of audio and image signals can better help the model locate the source target, rather than directly fusing entire audio signals with image information. Finally, to validate the effect of multi-frequency audio feature, after incorporating multi-frequency audio information into the model input, compared with using only single-frequency audio features, $\mathcal{M}_{\mathcal{J}}$ increased by 3.7. This indicates that after the addition of frequency channels, the audio features can adapt to more diverse sound source targets in mixed audio scenes.

Impact of S4. Following previous work [15], we investigated the impact of fine-tuning S4 model for MS3 task. The results are shown in Table 2. Currently, our method using the S4 pre-training weights has achieved the best performance on the MS3 subset. Because S4 has the same sound source categories and similar scenes as MS3, the pre-trained weights on the S4 naturally fit the MS3. Moreover, we also conducted related experiments in the S4 subset, as shown in Table 3. Our method does not diminish the detection effect of monophonic sound source targets. On the contrary, there is a certain improvement compared to the baseline. We believe this is due to AVMM and MCP, which allow the audio features to guide the image with more detail.

Attention Map Visualization. In order to qualitatively analyze the impact of frequency on the sound source target, we input high-frequency and low-frequency audio information respectively and visualize the attention map results. As shown in Fig. 6, in the first row, since the violin emits high-frequency audio, when low-frequency audio is input, the model will not focus on the target violin as the sound source, but the high-frequencyaudio features can find the violin. In the second row, guitar and ukulele produce low-frequency audio, which is obviously more inclined to low-frequency audio information in the visualized feature map. Therefore, when different objects emit sounds of different frequencies, choosing a more suitable frequency can improve the segmentation results of the model.

Table 2. Investigate the impact of different weight initialization strategies on the MS3 subset, as well as the influence of the prompt, AVMM, and multi-frequency audio on the model, respectively.

	From Scratch		P.T. on S4	
	$\mathcal{M}_\mathcal{J}$	$\mathcal{M}_\mathcal{F}$	$\mathcal{M}_\mathcal{J}$	$\mathcal{M}_\mathcal{F}$
baseline	54.0	64.5	58.4	68.6
add prompter	57.1	67.7	60.2	71.7
add AVMM	60.4	69.1	63.1	73.4
add multi-frequency audio	**64.1**	**73.9**	**67.3**	**76.96**

Table 3. Comparison with related work on the S4 subset of AVSbench

Methods	ResNet-50		PVT-v2	
	$\mathcal{M}_\mathcal{J}$	$\mathcal{M}_\mathcal{F}$	$\mathcal{M}_\mathcal{J}$	$\mathcal{M}_\mathcal{F}$
AVSBench [35]	72.8	84.8	78.7	87.9
AVS-BiGen [9]	74.1	85.4	81.7	90.4
BAVS [3]	78.0	85.3	82.0	88.6
COMBO [32]	81.7	90.1	84.7	91.9
ECMVCE [24]	76.3	86.5	81.7	90.1
DiffusionAVS [23]	75.8	86.9	81.4	90.2
AVSegFormer [7]	76.45	85.9	82.1	89.9
Ours	**78.9**	**87.4**	**83.3**	**90.8**

5 Conclusion

In this paper, we propose an audio-visual segmentation method for MS3, called multi-frequency fine-grained matching. We introduce multi-frequency audio information by expanding the frequency channel of audio features. In addition, our audio-visual matching module uses audio prompts to adaptively modulate visual features, which allows the precise alignment between visual and audio information. Our method provides a disparate target representation of different sound source targets, which improves the performance of AVS task in mixed audio scenarios. Experimental results show that our method has superior performance compared with the existing state-of-the-art methods.

References

1. Chen, H., Xie, W., Afouras, T., Nagrani, A., Vedaldi, A., Zisserman, A.: Localizing visual sounds the hard way. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 16867–16876 (2021)
2. Chen, L.C., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A.L.: Deeplab: semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs. IEEE Trans. Pattern Anal. Mach. Intell. **40**(4), 834–848 (2017)
3. Chen, T., et al.: Bootstrapping audio-visual segmentation by strengthening audio cues. arXiv preprint arXiv:2402.02327 (2024)
4. Cheng, B., Misra, I., Schwing, A.G., Kirillov, A., Girdhar, R.: Masked-attention mask transformer for universal image segmentation. arXiv (2021)
5. Dosovitskiy, A., et al.: An image is worth 16x16 words: transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
6. Gao, R., Grauman, K.: 2.5 D visual sound. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 324–333 (2019)
7. Gao, S., Chen, Z., Chen, G., Wang, W., Lu, T.: Avsegformer: audio-visual segmentation with transformer. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 38, pp. 12155–12163 (2024)
8. Gemmeke, J.F., et al.: Audio set: an ontology and human-labeled dataset for audio events. In: 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 776–780. IEEE (2017)
9. Hao, D., Mao, Y., He, B., Han, X., Dai, Y., Zhong, Y.: Improving audio-visual segmentation with bidirectional generation. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 38, pp. 2067–2075 (2024)
10. 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)
11. 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)
12. Huang, S., et al.: Discovering sounding objects by audio queries for audio visual segmentation. arXiv preprint arXiv:2309.09501 (2023)
13. Kirillov, A., et al.: Segment anything. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 4015–4026 (2023)
14. Li, F., et al.: Semantic-SAM: segment and recognize anything at any granularity. arXiv preprint arXiv:2307.04767 (2023)
15. Li, K., Yang, Z., Chen, L., Yang, Y., Xiao, J.: CATR: combinatorial-dependence audio-queried transformer for audio-visual video segmentation. In: Proceedings of the 31st ACM International Conference on Multimedia, pp. 1485–1494 (2023)
16. Lin, Y.B., Sung, Y.L., Lei, J., Bansal, M., Bertasius, G.: Vision transformers are parameter-efficient audio-visual learners. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 2299–2309 (2023)

17. Liu, C., et al.: BAVS: bootstrapping audio-visual segmentation by integrating foundation knowledge. arXiv preprint arXiv:2308.10175 (2023)
18. Liu, C., et al.: Audio-visual segmentation by exploring cross-modal mutual semantics. In: Proceedings of the 31st ACM International Conference on Multimedia, pp. 7590–7598 (2023)
19. Liu, L., Chang, J., Yu, B.X., Lin, L., Tian, Q., Chen, C.W.: Prompt-matched semantic segmentation. arXiv preprint arXiv:2208.10159 (2022)
20. Liu, P., Yuan, W., Fu, J., Jiang, Z., Hayashi, H., Neubig, G.: Pre-train, prompt, and predict: a systematic survey of prompting methods in natural language processing. ACM Comput. Surv. 1–35 (2023). https://doi.org/10.1145/3560815
21. Liu, Z., et al.: Swin transformer: hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) (2021)
22. Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3431–3440 (2015)
23. Mao, Y., Zhang, J., Xiang, M., Lv, Y., Zhong, Y., Dai, Y.: Contrastive conditional latent diffusion for audio-visual segmentation. arXiv preprint arXiv:2307.16579 (2023)
24. Mao, Y., Zhang, J., Xiang, M., Zhong, Y., Dai, Y.: Multimodal variational autoencoder based audio-visual segmentation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 954–965 (2023)
25. Morgado, P., Nvasconcelos, N., Langlois, T., Wang, O.: Self-supervised generation of spatial audio for 360 video. In: Advances in Neural Information Processing Systems, vol. 31 (2018)
26. Owens, A., Efros, A.A.: Audio-visual scene analysis with self-supervised multisensory features. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 631–648 (2018)
27. Qian, R., Hu, D., Dinkel, H., Wu, M., Xu, N., Lin, W.: Multiple sound sources localization from coarse to fine. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XX 16, pp. 292–308. Springer, Cham (2020)
28. Strudel, R., Garcia, R., Laptev, I., Schmid, C.: Segmenter: transformer for semantic segmentation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 7262–7272 (2021)
29. Wang, W., et al.: PVT V2: improved baselines with pyramid vision transformer. Comput. Visual Media **8**(3), 415–424 (2022)
30. Wang, Y., Liu, W., Li, G., Ding, J., Hu, D., Li, X.: Prompting segmentation with sound is generalizable audio-visual source localizer. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 38, pp. 5669–5677 (2024)
31. Xie, E., Wang, W., Yu, Z., Anandkumar, A., Alvarez, J.M., Luo, P.: Segformer: simple and efficient design for semantic segmentation with transformers. In: Neural Information Processing Systems (NeurIPS) (2021)
32. Yang, Q., et al.: Cooperation does matter: exploring multi-order bilateral relations for audio-visual segmentation. arXiv preprint arXiv:2312.06462 (2023)
33. Zhao, H., Shi, J., Qi, X., Wang, X., Jia, J.: Pyramid scene parsing network. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2881–2890 (2017)
34. 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)

35. Zhou, J., et al.: Audio–visual segmentation. In: Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXVII, pp. 386–403. Springer, Cham (2022)
36. Zhu, J., Lai, S., Chen, X., Wang, D., Lu, H.: Visual prompt multi-modal tracking. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 9516–9526 (2023)

Confidence-Guided Feature Alignment for Cloth-Changing Person Re-identification

Sirong Huang[1,2,3] and Huicheng Zheng[1,2,3](✉)

[1] School of Computer Science and Engineering, Sun Yat-Sen University, Guangzhou, China
huangsir5@mail2.sysu.edu.cn, zhenghch@mail.sysu.edu.cn

[2] Key Laboratory of Machine Intelligence and Advanced Computing, Ministry of Education, Guangzhou, China

[3] Guangdong Province Key Laboratory of Information Security Technology, Guangzhou, China

Abstract. Much progress has been made in the field of person re-identification, but changes in clothing have hindered the practical application of long-term person re-identification. Cloth-changing person re-identification (CC-ReID) aims to address this problem, with the main challenge being the extraction of discriminative features unrelated to clothing. Existing methods, which mainly focus on introducing clothing-irrelevant cues such as key points, contours, and 3D shapes, require additional modules for feature extraction, resulting in increased complexity and potential inaccuracies due to dependence on the performance of external models. Few studies directly use the original RGB images to make the model constantly focus on clothing-independent information. In this paper, we propose a Confidence-Guided Feature Alignment Network (CGFA) for CC-ReID. Specifically, we design a confidence module that automatically learns to make confidence adjustments to fine-grained information, prompting the model to mine clothing-independent discriminative features without introducing other modal cues. By transferring knowledge, we encourage the model to learn discriminative identity features that are independent of clothing bias. As a result, the confidence module can be removed during the inference phase. The proposed simple but efficient method uses only RGB modality without additional cues, and can serve as a powerful baseline for CC-ReID to drive future research. Extensive experiments on the CC-ReID datasets demonstrate the effectiveness of the proposed method, which achieves state-of-the-art performance.

Keywords: Person re-identification · Cloth-changing · Transformer · Confidence-guided

A. Antonacopoulos et al. (Eds.): ICPR 2024, LNCS 15329, pp. 51–64, 2025.
https://doi.org/10.1007/978-3-031-78110-0_4

1 Introduction

Person re-identification (ReID) refers to matching the same pedestrian within distributed camera systems. Most current ReID researches are based on the assumption that each pedestrian appears under the cameras for a relatively short period of time and that the pedestrian's clothing does not change [9,22,26]. However, the problem of clothing changes is unavoidable in practice. Take two examples, one is when an elder or a child is lost and the photo provided by the family does not match the clothing worn by the pedestrian at the time of loss. Another example is that suspects often change their clothing to avoid recognition and tracking, resulting in significant changes to their visual appearance characteristics. This leads to a more complex and challenging task, namely, Cloth-Changing Person Re-identification (CC-ReID).

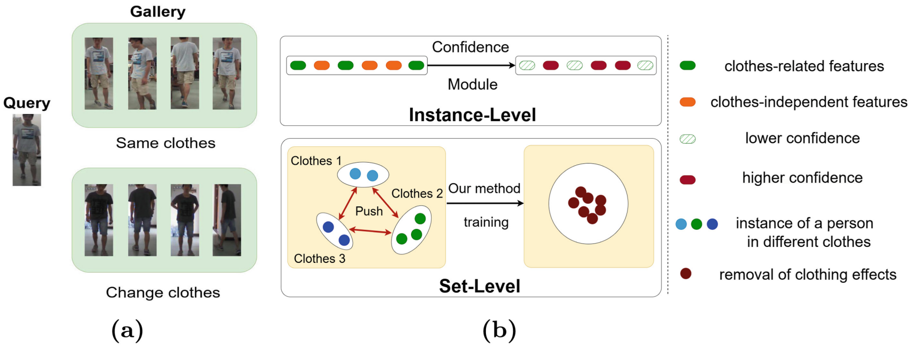


Fig. 1. (a) Difference between the general person ReID and CC-ReID tasks, where the change of clothes introduces significant intra-class differences. (b) The proposed method reduces the confidence score of clothing-related features, thereby reducing cross-clothing variation and focusing more on clothing-independent features. Importantly, this approach obviates the need for additional cues.

Clothing covers a large part of the human body, so there are huge differences in the appearance of the same person wearing different clothes, as shown in Fig. 1(a). The difference in colour and texture information leads to an increase in intra-class distance for the same pedestrian after changing clothes, which is the main challenges of CC-ReID. Existing methods [4,14,17,24,31] mainly focus on the acquisition of clothing-independent cues such as key points, contours, 3D shapes, gait, etc. While these clothing-independent cues are helpful, the additional computational cost of training and even inference to obtain them is high. In addition, we note that these methods ignore how the original RGB images can be used directly to improve the similarity of same pedestrians across clothes from a confidence perspective, which is useful for CC-ReID.

As the same pedestrian wears different clothes, the large difference in the clothing part will eventually increase the intra-class distance, as shown in

Fig. 1(b). Therefore, we readjust the contribution of the different parts to the overall features using the confidence module, i.e. we decrease the confidence score of the clothes-related features and increase the confidence score of discriminative features, thus increasing the similarity of same pedestrians with different clothes. Specifically, our confidence module uses fine-grained alignment information between pedestrians and uses the similarity of the aligned parts as the confidence score. In this way, differences in clothing lead to a natural decrease in the confidence score of the parts associated with the clothing and an increase in the confidence score of invariant features. We perform knowledge migration after correcting for clothing bias, so the confidence module can be omitted during the inference phase.

The contributions of this paper can be summarised as follows:

- We design the confidence module specifically for CC-ReID, which aims to reduce the confidence of clothing-related features and increase the confidence of clothing-independent features, thus reducing the intra-class distance and increasing the similarity of the same person.
- We propose a simple yet effective method for CC-ReID without additional cues. The method can be used as a transformer-based CC-ReID baseline.
- We conduct extensive experiments on three CC-ReID datasets to demonstrate the effectiveness of the proposed method.

2 Related Work

Person ReID. The goal of ReID is to match the same pedestrian in different camera views, a task that faces challenges such as complex backgrounds, occlusions, changing lighting conditions, and different viewpoints. Convolutional Neural Networks (CNNs) have long been the mainstay of ReID research [1], which can be divided into representation-based learning and metric-based learning according to training loss. This has changed with the advent of Vision Transformer [7,27], a breakthrough in computer vision that has demonstrated excellent performance in a wide range of vision tasks, including ReID. Transformer is a network architecture dispensing with recurrence and convolutions that relies entirely on attention mechanisms to model the global dependencies between inputs and outputs [33]. Some research in ReID has significantly improved performance by replacing CNNs with transformer as feature extractors [3,5,6,13,35]. However, none of these methods consider the issue of clothing variations, making them ineffective for CC-ReID.

Cloth-Changing Person ReID. As the field of ReID has developed, a number of researchers have highlighted the fact that existing ReID methods rely heavily on the clothing characteristics of pedestrians [16,20,24,28,29]. Most of the existing methods use some additional modules to extract clothing-independent information to guide the model training. [31] uses contour sketches from human images for CC-ReID, but these sketches lack human body details. [17] tackles CC-ReID by leveraging gait as supplementary data, but obtaining effective gait information for pedestrians facing front or wearing skirts is difficult. Methods such as FSAM [14], SPS [25] and M2Net [21] use existing parsing networks to obtain contour images for training, these methods discard all the colour information in the original RGB image in the contour processing module, but some of the colour information helps to re-identify the person. All of these methods are prone to estimation errors introduced by the extractor, resulting in increased computational cost. In contrast, our approach focuses on directly using the features from the original RGB image to mitigate clothing interference, thus eliminating the need for an additional extraction module and avoiding dependence on the performance of other models, such as segmentation models and gait extraction. Our method can integrate with existing approaches, serving as an approach to provide more robust features.

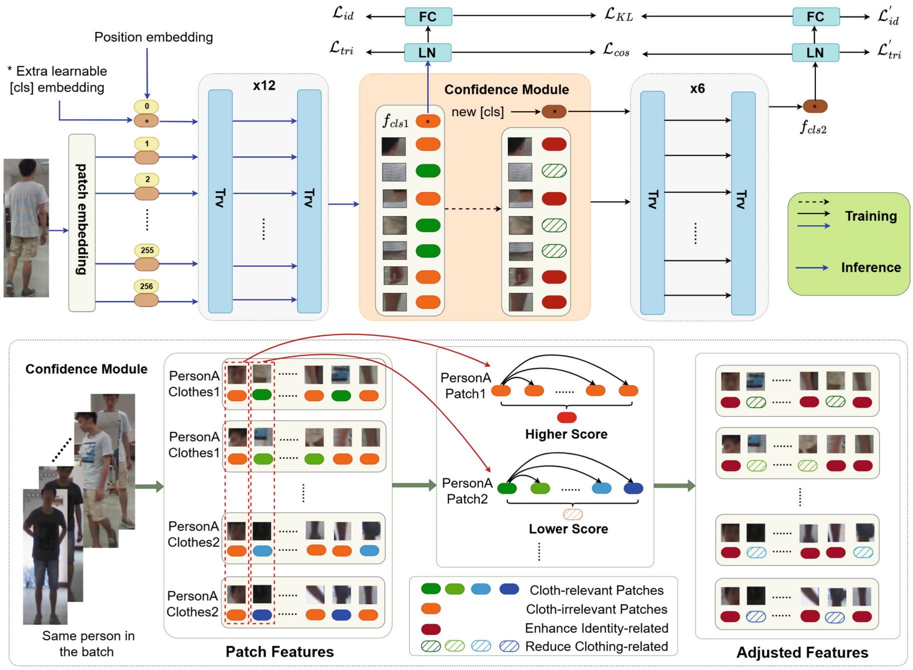


Fig. 2. Overview of Confidence-Guided Feature Alignment Network. Our network mines more clothing-independent information by decreasing the confidence score of clothing-related features and increasing the confidence score of clothing-independent features. The confidence module uses cosine similarity to calculate confidence score of aligned patches from the same pedestrian.

3 Method

In this section, we present the proposed Confidence-Guided Feature Alignment Network, outlined in the framework shown in Fig. 2. Through our CGFA approach, we diminish the confidence score associated with clothing-related features while enhancing those corresponding to clothing-independent features. This process helps the model to learn discriminative features that are unrelated to clothing. Additionally, we employ cosine similarity loss and Kullback Leibler (KL) Divergence [18] to encourage the model to learn consistent feature representations before and after the confidence module, thereby obviating the necessity for confidence module during the inference phase.

3.1 Overview

Let $x_j \in \mathbb{R}^{H \times W \times C}$ be the j-th pedestrian image, where W, H and C denote the width, height and number of channels of the image, respectively. We partition x_j into a sequence of patches $\{p_1, p_2, \ldots, p_N\}$, where N represents the length of fixed-size sequences, defined as $N = H \times W/P^2$, with P denoting the patch size. Then we embed each patch with linear projection $\varepsilon(\cdot)$ into D dimensions. A learnable [CLS] token is added to the token sequence and positional embedding is applied to each patch. The final input can be described as:

$$Z = \{x_{cls}, \varepsilon(p_1), \varepsilon(p_2), \ldots, \varepsilon(p_N)\} + \mathcal{P} \tag{1}$$

where $x_{cls} \in \mathbb{R}^{1 \times D}$ is a learnable [CLS] token, $\varepsilon(p_i) \in \mathbb{R}^{1 \times D}$ is i-th patch embeddings, $\mathcal{P} \in \mathbb{R}^{(N+1) \times D}$ is position embeddings. Z is then fed into the TrV blocks [8] of L layers to capture the dependencies between patches. The TrV block is an improved Vision Transformer structure. We use EVA02-large [8] as the backbone, which has 24 layers of TrV blocks. The extracted f_{cls1} is used as a feature representation of the global image. To further improve the cross-clothing retrieval ability of the model, we automatically adjust the confidence score of the extracted features $\{f_{p_i}|_{i=1}^{N}\}$ using the confidence module. The confidence module reduces the confidence score for clothing interference, the details of which are described in the next subsection.

3.2 Confidence Module

The confidence module is used to adjust the confidence score of patches. Although the design of the transformer can simulate the dependencies between patches to obtain a global feature representation, it lacks the consideration of pedestrians changing clothes. Therefore, we obtain the confidence score of patches by batch computing the average similarity of aligned patches from the same pedestrian, as shown in Fig. 2, which can be expressed for x_j as

$$S_{p_i}^{j} = \frac{1}{K-1} \sum_{k=1, k \neq j}^{K} \delta(f_{p_i}^{j}, f_{p_i}^{k}) \tag{2}$$

where K denotes the number of images belonging to this ID. $\delta(\cdot, \cdot)$ denotes cosine similarity function, which is defined as $\delta(u, v) = \frac{u}{||u||} \cdot \frac{v}{||v||}$. Then the confidence score of the clothing related patches will decrease due to the differences after the clothing change, whereas the confidence score of the clothing unrelated patches will remain high due to the high similarity. The confidence score of the corresponding patch is used to correct the original patch. The new token sequence is then formed by attaching a new [CLS] token, as follows,

$$Z^{'} = \{x^{'}_{cls}, S^{j}_{p_1} f^{j}_{p_1}, S^{j}_{p_2} f^{j}_{p_2}, \ldots, S^{j}_{p_N} f^{j}_{p_N}\} \tag{3}$$

then $Z^{'}$ is fed into the TrV blocks of $L^{'}$ layers. The dependencies between patches are then modelled to obtain f_{cls2}, which ultimately guides f_{cls1}:

$$\mathcal{L}_{cos} = 1 - \delta(LN(f_{cls1}), LN(f_{cls2})) \tag{4}$$

Here, LN refers to the layer normalization layer and f_{cls2} represents the features after correcting for clothing bias. In addition, we adopt the KL Divergence as in mutual learning [36] to further allow f_{cls1} to perceive the eliminated clothing bias in f_{cls2}:

$$Q^1 = exp(FC(LN(f_{cls1}))), Q^2 = exp(FC(LN(f_{cls2}))) \tag{5}$$

$$D_{KL}(Q^1||Q^2) = \frac{1}{n}\sum_{j=1}^{n}\sum_{m=1}^{M} Q^1_{j,m} log \frac{Q^1_{j,m}}{Q^2_{j,m}} \tag{6}$$

where n is the number of images in a mini-batch and M is the number of identities in a dataset. To be noticed, due to the asymmetry of KL Divergence, we also compute $D_{KL}(Q^2||Q^1)$. The total KL Divergence can be formulated as:

$$\mathcal{L}_{KL} = D_{KL}(Q^1||Q^2) + D_{KL}(Q^2||Q^1) \tag{7}$$

Instructing f_{cls1} with f_{cls2}, which eliminates the clothing bias, makes it possible to obtain discriminative cloth-irrelevant features without the need for an additional module in inference.

3.3 Loss Function

We introduce an identity loss to extract the identity information. The identity loss $\mathcal{L}_{id}$ is a cross-entropy loss, which can be denoted as:

$$\mathcal{L}_{id} = -\frac{1}{n}\sum_{j=1}^{n} log(p(y_j|x_j)) \tag{8}$$

where y_j is the label of image x_j. In addition, we use a triplet loss to minimize the distance between similar images and maximize the distance between dissimilar images, which can be expressed as:

$$\mathcal{L}_{tri} = max\{margin + D(x_{jA}, x_{jP}) - D(x_{jA}, x_{jN}), 0\} \tag{9}$$

where $D(\cdot)$ is the squared Euclidean distance in the embedding space. x_{jA}, x_{jP}, x_{jN} are anchor images, positive samples, and negative samples, respectively.

We mix the loss functions into the following form according to the type of loss functions. We assign equal importance to label smooth classification loss $\mathcal{L}_{id}$ and triplet loss $\mathcal{L}_{tri}$. Then we use λ_{cos} and λ_{KL} to control the weights of $\mathcal{L}_{cos}$ and $\mathcal{L}_{KL}$, respectively. In summary, the overall loss is as follows:

$$\mathcal{L} = \lambda_1 \mathcal{L}_{id} + \lambda_2 \mathcal{L}_{tri} + \lambda_1 \mathcal{L}_{id}^{'} + \lambda_2 \mathcal{L}_{tri}^{'} + \lambda_{cos} \mathcal{L}_{cos} + \lambda_{KL} \mathcal{L}_{KL} \tag{10}$$

Referring to Fig. 2, $\mathcal{L}^{'}$ represents the supervised loss after correcting for clothing bias. We optimize the entire network in an end-to-end network by minimizing the overall loss function.

4 Experiment

4.1 Datasets

We mainly evaluate our method on three cloth-changing ReID datasets: PRCC [31], LTCC [24] and VC-Clothes [28].

PRCC dataset contains images from three cameras (A, B, and C), where A and B capture images of the same clothing in different scenarios, A and C capture images of different clothing. The dataset contains 221 identities and 33,698 images, providing a comprehensive training and testing set, with corresponding contour sketches provided for each image.

LTCC is another large CC-ReID dataset, captured by 12 cameras. The dataset spans a period of two months and collects 162 identities and 15,138 images. The dataset is divided into two subsets: one subset contains 91 individuals wearing different outfits, consisting of 415 outfits and 14,756 images; the other subset contains 61 individuals, with each person wearing the same outfit in all their images, totaling 2,382 images.

VC-Clothes is a synthetic dataset rendered by the GTA5 game engine with 512 identities, 4 cameras and an average of 9 images per scene per identity, for a total of 19,060 images. There are 9,449 training images, 1,020 query images and 8,591 gallery images.

4.2 Evaluation Settings and Protocol

Following the conventions of the ReID community, we evaluate all methods using rank-K and the mean Average Precision (mAP). Following [10], three test settings are defined as (1) **general setting** (General): both clothes-changing and clothes-consistent gallery samples are used to calculate the accuracies, (2) **cloth-changing setting** (CC): only clothes-changing gallery samples are used to calculate the accuracies, and (3) **same-clothes setting** (SC): only clothes-consistent gallery samples are used to calculate the accuracies. We report Rank-1 accuracy and mAP for all datasets for evaluation.

4.3 Implementation Details

We use the pre-trained EVA02-large model as our backbone network. In the training phase, we extract f_{cls1} and f_{cls2} using 24 and 12 TrV blocks respectively. In the testing phase, only the features extracted from the first 24 TrV blocks are used, without forward propagating the remaining 12 TrV blocks. Random cropping and erasing [37] are used for data augmentation. The model is trained for 60 epochs using the SGD optimizer. The warmup learning rate is initially set to $7.8125e^{-7}$. The learning rate is initially set to $2e^{-5}$ and divided by 100 at 40 and 60 epochs. The input images are resized to 224×224 for all datasets. The batchsize is set to 8, where each batch includes two different persons with 4 images for each person. As for the hyperparameters, both λ_1 and λ_2 in Eq. 10 are set to 1. λ_{cos} and λ_{KL} are set to 0.8 and 0.5, respectively.

Table 1. Comparison of Rank-1 accuracy(%) and mAP(%) with the state-of-the-art methods on PRCC and LTCC, where "sketch", "pose", "sil.", "parsing", "Gait", "clothes ID" and "3D" denote contour sketches, keypoints, silhouettes, human parsing, gait information, clothes labels and 3D shape information. Bold and underlined numbers are the best and second best scores, respectively.

Method	Auxiliary Information	PRCC				LTCC			
		CC		SC		CC		General	
		Rank-1	mAP	Rank-1	mAP	Rank-1	mAP	Rank-1	mAP
IANet (CVPR 19) [15]	none	46.3	45.9	99.4	98.3	25.0	12.6	63.7	31.0
ISP (ECCV 20) [38]	none	36.6	-	92.8	-	27.8	11.9	66.3	29.6
SPT+ASE (TPAMI 19) [31]	sketch	34.4	-	64.2	-	-	-	-	-
FSAM (CVPR 21) [14]	pose+sil.+parsing	54.5	-	98.8	-	38.5	16.2	73.2	35.4
GI-ReID (CVPR 22) [17]	parsing+sil.+Gait	-	37.5	-	-	23.7	10.4	63.2	29.4
UCAD (IJCAI 22) [30]	sil.	45.3	-	96.5	-	32.5	15.1	74.4	34.8
ViT-VIBE (WACV 22) [2]	3D	47.0	-	99.7	-	-	-	71.4	35.8
CAL (CVPR 22) [10]	clothes ID	55.2	55.8	**100.0**	99.8	40.1	18.0	74.2	40.8
CCFA (CVPR 23) [12]	clothes ID	61.2	58.4	99.6	98.7	45.3	22.1	75.8	42.5
AIM (CVPR 23) [32]	clothes ID	57.9	58.3	**100.0**	**99.9**	40.6	19.1	76.3	41.1
SCNet (ACMMM 23) [11]	parsing	61.3	**59.9**	**100.0**	97.8	47.5	25.5	76.3	43.6
CGFA (Ours)	none	**64.6**	59.7	**100.0**	98.8	**49.0**	**27.1**	**83.0**	**49.6**

Table 2. Comparison of Rank-1 accuracy(%) and mAP(%) with the state-of-the-art methods in cloth-changing setting on VC-Clothes. Bold and underlined numbers are the best and second best scores, respectively.

Method	FSAM [14]	GI-ReID [17]	UCAD [30]	CAL [10]	MVGD [34]	CACC [19]	SCNet [11]	CGFA (Ours)
Rank-1	78.6	64.5	82.4	81.4	82.6	85.0	90.1	**94.3**
mAP	78.9	57.8	73.8	81.7	78.4	81.2	84.4	**88.0**

4.4 Comparison with the State-of-the-Art Methods

We compare the proposed CGFA method with the state-of-the-art methods. Two general ReID methods (i.e., IANet [15] and ISP [38]) and nine CC-ReID methods are included. The results are reported in Table 1 and Table 2. Notably, the majority of CC-ReID methodologies outperform general ReID approaches, which can be attributed to the integration of auxiliary modules or optimization functionalities aimed at mitigating the impact of clothing variations. In contrast to these methods, the proposed CGFA approach can obtain features that are more robust to appearance changes by correcting for clothing bias through the confidence module to ensure that the model focuses on clothing-independent information. As a result, our method achieves significant performance improvement over the SOTA method under the cloth-changing setting. Specifically, we achieve 64.6% Rank-1 and 59.7% mAP on the PRCC dataset. In the cloth-changing setting of the LTCC dataset, we have 1.5% improvement in Rank-1 and 1.6% improvement in mAP over the SOTA method SCNet [11]. The results show that even under complex scenarios (changing resolution, illumination, viewpoint, et al.), CGFA can still capture the existing discriminative features rather than being misled by the noticeable clothing bias. We also achieve the best performance on the VC-Clothes in Rank-1 and mAP, 94.3% and 88.0%, respectively. While tailored for CC-ReID, our approach demonstrates competitive performance across general and same-clothes setting. Notably, on the LTCC dataset, it attains the highest performance, achieving 83.0% at Rank-1 and 49.6% for mAP. These results underscore the efficacy of CGFA in thoroughly exploring fine-grained identity-related information.

Table 3. Ablation studies of CGFA in clothe-changing setting on PRCC and LTCC, where the baseline contains 24 layers of TrV blocks, the baseline* contains 36 layers of TrV blocks, and CM stands for the confidence module.

Method	PRCC		LTCC	
	Rank-1	mAP	Rank-1	mAP
baseline	62.4	57.4	45.9	23.8
baseline*	61.0	56.6	44.1	23.2
CM+$\mathcal{L}_{cos}$	63.4	59.5	47.7	26.5
CM+$\mathcal{L}_{KL}$	63.5	58.9	47.4	26.8
CM+$\mathcal{L}_{cos}$+$\mathcal{L}_{KL}$ (CGFA)	**64.6**	**59.7**	**49.0**	**27.1**

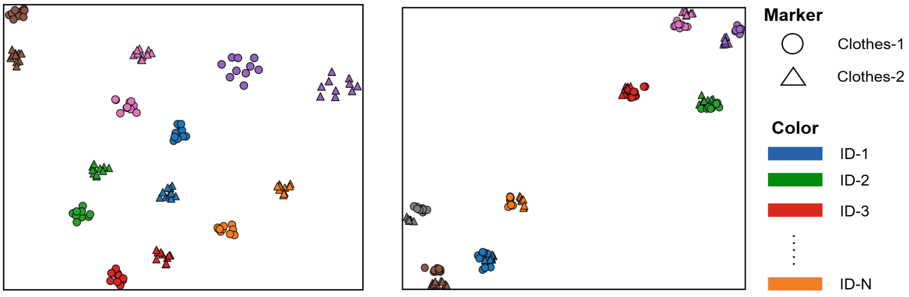
Table 4. Ablation studies on the different numbers of TrV blocks after the confidence module for LTCC in clothing-changing setting.

Layer	LTCC	
	Rank-1	mAP
6	48.0	25.4
12	**49.0**	**27.1**
18	48.5	26.0
24	46.7	24.3

4.5 Ablation Study

As shown in Table 3 and Table 4, we perform comprehensive ablation studies. In this part, we explore the role of the confidence module on the model's ability to learn clothing-independent features and the effect of different numbers of TrV blocks.

Effectiveness of the Confidence Module. We use EVA02-large [8] encoded by position embedding as the baseline. Baseline* is the version with 36 layers of TrV blocks, which is used to illustrate that it is the confidence module rather than the increased number of TrV blocks that brings about the improvement in our method. We also test the effect of cosine similarity loss and KL loss on knowledge transfer learning. We conduct experiments for both PRCC and LTCC in cloth-changing setting. The results are summarised in Table 3. The results of baseline* decrease compared with the baseline, suggesting that it is not the case that the higher the number of TrV blocks, the better the performance. It can be observed that both cosine similarity loss and KL loss contribute to knowledge migration. The best performance is achieved when both are used simultaneously, outperforming the baseline by 2.2%/3.1% in Rank-1 and 2.3%/3.3% in mAP on PRCC and LTCC. The results suggest that the confidence module can force



(a) Baseline (b) Ours

Fig. 3. The t-SNE visualization of features on the PRCC dataset. We randomly choose 8 identities, corresponding to specific colors in the figure.

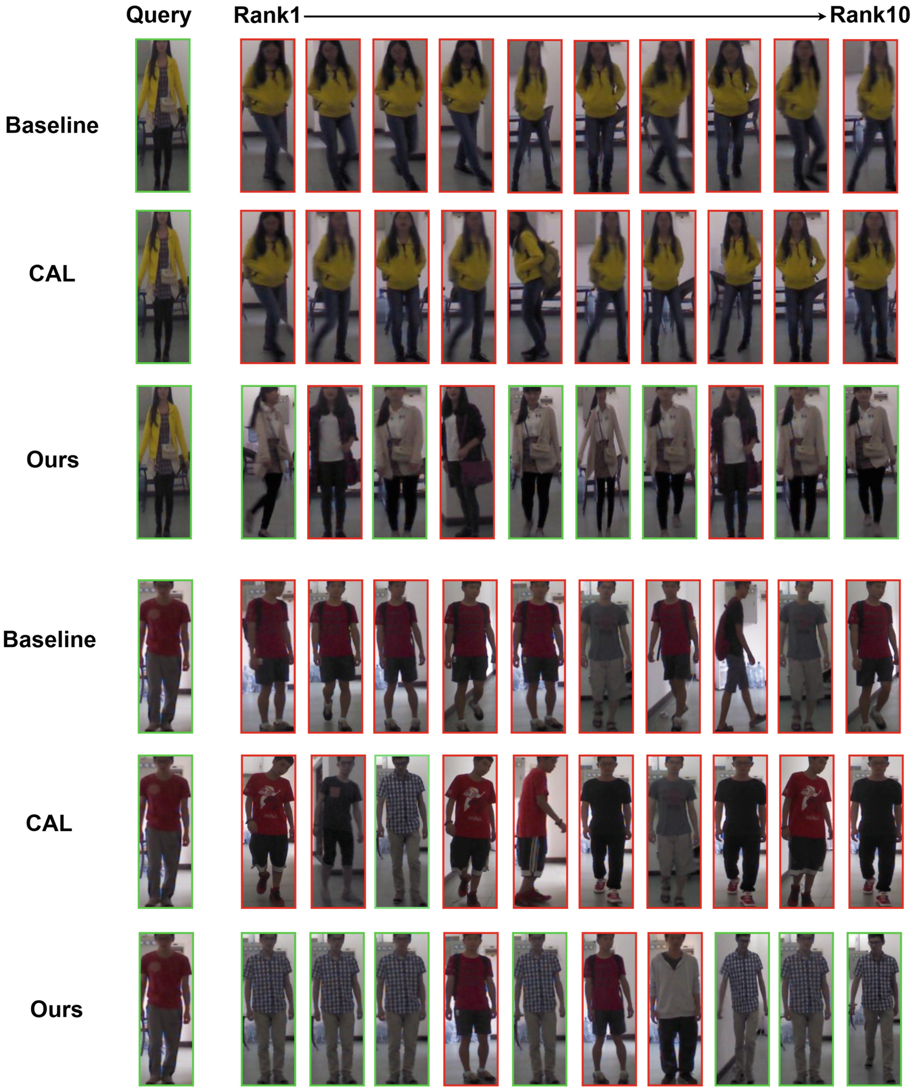


Fig. 4. Visualisation of the top 10 rankings for the baseline, CAL [10], and our CGFA on the PRCC dataset. Images in green and red boxes correspond to positive and negative results, respectively.

the model to reduce the confidence score of clothing-related features and learn clothing-insensitive features.

Number of Trv Blocks. In this subsection, we discuss the effect of the number of TrV blocks after the confidence module. We empirically try several layer combinations and summarise the experimental results for LTCC clothe-changing setting in Table 4. We can observe that with only six TrV blocks the Rank-1 and mAP are 2.1% and 1.6% higher than the baseline, respectively, which shows the effectiveness of the confidence module. The combination of 24 and 12 with the best effect is finally chosen for our network structure.

4.6 Visualization and Analysis

Visualization of Feature Distribution. To investigate the distribution of the features learned by the model, we perform t-SNE [23] visualisation experiments on the testing set of PRCC, comparing the distribution of features in the latent space for the baseline and the proposed CGFA. As shown in Fig. 3, the baseline shows large intra-class differences for different clothes with the same ID. In contrast, the CGFA learns features with better intra-class compactness. This suggests that our method can effectively learn identity-related discriminative features and mitigate the interference caused by clothing variations.

Visualization of Retrieval Results. We visually compare the baseline, CAL [10], and the proposed CGFA on the PRCC dataset. Figure 4 shows the top 10 retrieval results under the cloth-change setting. The green and red boxes indicate the positive and negative retrieval results, respectively. The results show that the baseline and CAL are affected by the colour and texture of clothing and incorrectly match people with similar clothing. On the other hand, our method employs the proposed confidence module and mutual learning to effectively counteract the clothing interference and shows better retrieval quality.

5 Conclusion

In this paper, we propose a confidence module to adjust the confidence score of features to deal with the cloth-changing ReID problem. This confidence module reduces the confidence score of clothing-related features and increases the confidence score of clothing-independent features to reduce intra-class variations when clothes change. Our method does not contain additional network branches or cues at the inference stage, so the computational cost is lower than other methods. Experimental results demonstrate the effectiveness of the proposed method.

Acknowledgements. This work was supported in part by the National Natural Science Foundation of China under Grant 61976231 and in part by the Guangdong Basic and Applied Basic Research Foundation under Grant 2023A1515012853 and Grant 2019A1515011869.

References

1. Ahmed, E., Jones, M., Marks, T.K.: An improved deep learning architecture for person re-identification. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 3908–3916 (2015)

2. Bansal, V., Foresti, G.L., Martinel, N.: Cloth-changing person re-identification with self-attention. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pp. 602–610 (2022)
3. Cao, J., et al.: PSTR: end-to-end one-step person search with transformers. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 9458–9467 (2022)
4. Chen, J., et al.: Learning 3D shape feature for texture-insensitive person re-identification. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 8146–8155 (2021)
5. Chen, S., Ye, M., Du, B.: Rotation invariant transformer for recognizing object in UAVs. In: Proceedings of the ACM International Conference on Multimedia, pp. 2565–2574 (2022)
6. Comandur, B.: Sports re-id: improving re-identification of players in broadcast videos of team sports. arXiv preprint arXiv:2206.02373 (2022)
7. Dosovitskiy, A., et al.: An image is worth 16x16 words: transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
8. Fang, Y., Sun, Q., Wang, X., Huang, T., Wang, X., Cao, Y.: EVA-02: a visual representation for neon genesis. arXiv preprint arXiv:2303.11331 (2023)
9. Gong, S., Cristani, M., Loy, C.C., Hospedales, T.M.: The re-identification challenge. Person re-identification, pp. 1–20 (2014)
10. Gu, X., Chang, H., Ma, B., Bai, S., Shan, S., Chen, X.: Clothes-changing person re-identification with RGB modality only. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 1060–1069 (2022)
11. Guo, P., Liu, H., Wu, J., Wang, G., Wang, T.: Semantic-aware consistency network for cloth-changing person re-identification. In: Proceedings of the ACM International Conference on Multimedia, pp. 8730–8739 (2023)
12. Han, K., Gong, S., Huang, Y., Wang, L., Tan, T.: Clothing-change feature augmentation for person re-identification. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 22066–22075 (2023)
13. He, S., Luo, H., Wang, P., Wang, F., Li, H., Jiang, W.: Transreid: transformer-based object re-identification. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 15013–15022 (2021)
14. Hong, P., Wu, T., Wu, A., Han, X., Zheng, W.S.: Fine-grained shape-appearance mutual learning for cloth-changing person re-identification. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 10513–10522 (2021)
15. Hou, R., Ma, B., Chang, H., Gu, X., Shan, S., Chen, X.: Interaction-and-aggregation network for person re-identification. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 9317–9326 (2019)
16. Huang, Y., Wu, Q., Xu, J., Zhong, Y.: Celebrities-reid: a benchmark for clothes variation in long-term person re-identification. In: Proceedings of the International Joint Conference on Neural Networks, pp. 1–8 (2019)
17. Jin, X., et al.: Cloth-changing person re-identification from a single image with gait prediction and regularization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 14278–14287 (2022)
18. Joyce, J.M.: Kullback-Leibler Divergence. Springer, Heidelberg (2011)
19. Li, X., Liu, B., Lu, Y., Chu, Q., Yu, N.: Cloth-aware center cluster loss for cloth-changing person re-identification. In: Proceedings of the Chinese Conference on Pattern Recognition and Computer Vision, pp. 527–539. Springer, Cham (2022)
20. Li, Y.J., Luo, Z., Weng, X., Kitani, K.M.: Learning shape representations for clothing variations in person re-identification. arXiv preprint arXiv:2003.07340 (2020)

21. Liu, M., Ma, Z., Li, T., Jiang, Y., Wang, K.: Long-term person re-identification with dramatic appearance change: algorithm and benchmark. In: Proceedings of the ACM International Conference on Multimedia, pp. 6406–6415 (2022)
22. Luo, H., Gu, Y., Liao, X., Lai, S., Jiang, W.: Bag of tricks and a strong baseline for deep person re-identification. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (2019)
23. Van der Maaten, L., Hinton, G.: Visualizing data using t-SNE. J. Mach. Learn. Res. **9**(11) (2008)
24. Qian, X., et al.: Long-term cloth-changing person re-identification. In: Proceedings of the Asian Conference on Computer Vision (2020)
25. Shu, X., Li, G., Wang, X., Ruan, W., Tian, Q.: Semantic-guided pixel sampling for cloth-changing person re-identification. IEEE Signal Process. Lett. **28**, 1365–1369 (2021)
26. Sun, Y., Zheng, L., Yang, Y., Tian, Q., Wang, S.: Beyond part models: person retrieval with refined part pooling (and a strong convolutional baseline). In: Proceedings of the European Conference on Computer Vision, pp. 480–496 (2018)
27. Vaswani, A., et al.: Attention is all you need. In: Advances in Neural Information Processing Systems, vol. 30 (2017)
28. Wan, F., Wu, Y., Qian, X., Chen, Y., Fu, Y.: When person re-identification meets changing clothes. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, pp. 830–831 (2020)
29. Wang, K., Ma, Z., Chen, S., Yang, J., Zhou, K., Li, T.: A benchmark for clothes variation in person re-identification. Int. J. Intell. Syst. **35**(12), 1881–1898 (2020)
30. Yan, Y., et al.: Weakening the influence of clothing: universal clothing attribute disentanglement for person re-identification. In: Proceedings of the International Joint Conference on Artificial Intelligence, pp. 1523–1529 (2022)
31. Yang, Q., Wu, A., Zheng, W.S.: Person re-identification by contour sketch under moderate clothing change. IEEE Trans. Pattern Anal. Mach. Intell. **43**(6), 2029–2046 (2019)
32. Yang, Z., Lin, M., Zhong, X., Wu, Y., Wang, Z.: Good is bad: causality inspired cloth-debiasing for cloth-changing person re-identification. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 1472–1481 (2023)
33. Ye, M., Shen, J., Lin, G., Xiang, T., Shao, L., Hoi, S.C.: Deep learning for person re-identification: a survey and outlook. IEEE Trans. Pattern Anal. Mach. Intell. **44**(6), 2872–2893 (2021)
34. Yu, H., Liu, B., Lu, Y., Chu, Q., Yu, N.: Multi-view geometry distillation for cloth-changing person REID. In: Proceedings of the Chinese Conference on Pattern Recognition and Computer Vision, pp. 29–41. Springer, Cham (2022)
35. Zhang, B., Liang, Y., Du, M.: Interlaced perception for person re-identification based on swin transformer. In: Proceedings of the International Conference on Image, Vision and Computing, pp. 24–30. IEEE (2022)
36. Zhang, Y., Xiang, T., Hospedales, T.M., Lu, H.: Deep mutual learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 4320–4328 (2018)
37. Zhong, Z., Zheng, L., Kang, G., Li, S., Yang, Y.: Random erasing data augmentation. In: Proceedings of the AAAI Conference on Artificial Intelligence, pp. 13001–13008 (2020)
38. Zhu, K., Guo, H., Liu, Z., Tang, M., Wang, J.: Identity-guided human semantic parsing for person re-identification. In: Proceedings of the European Conference on Computer Vision, pp. 346–363. Springer, Cham (2020)

Fair Latent Representation Learning with Adaptive Reweighing

Puspita Majumdar(✉), Raghav Sharma, Rohit Bhattacharya, and Balraj Prajesh

AI Garage, Mastercard, Mumbai, India
{puspita.majumdar,raghav.sharma,rohit.bhattacharya, balraj.prajesh}@mastercard.com

Abstract. Artificial intelligence (AI) systems play significant roles in decision making processes, yet concerns persist about potential biases that can lead to unfair outcomes. These biases can arise from two main sources: imbalanced training data distribution and correlations between sensitive attributes (such as race and gender) and the target variable. Conventional model training methods penalize performance for underrepresented groups, resulting in biased outcomes. Further, they may capture features related to sensitive attributes during training, thus exacerbating bias. Biased predictions can have detrimental consequences. To address these concerns, in this research, we propose a novel method for bias mitigation. The proposed method aims to learn fair latent representations by emphasizing task relevant features while suppressing those linked to sensitive attributes. Additionally, we employ an adaptive reweighing technique to balance target class labels during training. The proposed method is evaluated on prominent benchmark datasets and compared with existing algorithms to demonstrate its effectiveness toward bias mitigation.

Keywords: AI fairness · Bias Mitigation · Fairness Aware Machine Learning · Responsible AI · Imbalanced Data · Ethical AI Development

1 Introduction

With technological advancement, Artificial Intelligence (AI) based systems are widely used in several applications to support decisions and make important predictions. However, on multiple occasions, AI systems have shown biased behavior, favoring certain groups of people over others [25]. This bias can manifest in various ways, such as word embeddings trained on large text corpora exhibiting historical bias. For example, a study found that word embeddings trained on Google News articles reflect and reinforce gender-based stereotypes in society. The word "man" is more closely associated with "computer programmer" than "woman", while the word "woman" is more closely associated with "homemaker" than "man" [3]. In another instance, ProPublica investigated predictive policing algorithms that are used to predict recidivism (the likelihood that someone

A. Antonacopoulos et al. (Eds.): ICPR 2024, LNCS 15329, pp. 65–80, 2025.
https://doi.org/10.1007/978-3-031-78110-0_5

will commit another crime) and found that they exhibit measurement bias. This leads to black defendants getting harsher sentences than white defendants for the same crime [27]. Such instances highlight that biased predictions can have multiple adverse effects on certain groups of people. Thus, bias in AI systems is a major concern regarding the ethical, social, and legal implications of developing and deploying AI systems.

There are several factors that cause AI systems to inherit, amplify, or create bias. One major factor is the quality and representativeness of the training data [30]. To overcome these issues, researchers have proposed various approaches to improve the data quality, such as oversampling underrepresented subgroups [21] or using data augmentation techniques to generate fair training data for bias mitigation [34]. Another major reason for bias in model prediction is the correlation of the sensitive attributes (e.g., race and gender) with the target variable [17]. In the conventional model training approach, the model is trained for a downstream task by automatically identifying and learning features that maximize the overall performance. However, in the learning process, the model may learn features related to the sensitive attributes, leading to unfair outcomes.

To address these challenges, we propose an algorithm that learns features related to the downstream task while ignoring the features related to the sensitive attribute to obtain a fair feature representation. We further leverage the concept of adaptive reweighing to deal with the imbalance in training data distribution and improve the overall performance of the proposed model. Multiple experiments have been performed on three benchmark datasets: *ADULT* [1], *COMPAS* [27] and *German-Credit-Risk* [13] to demonstrate the effectiveness of the proposed method for bias mitigation.

2 Previous Works

Researchers have proposed various techniques throughout the three steps of the machine learning pipeline (pre-processing, in-processing, and post-processing) to address bias and fairness in models [16]. Pre-processing methods focus on adjusting the training data, in-processing methods directly incorporate the fairness considerations into the model design itself, and post-processing methods address bias in the final output to achieve the goal of fairness. Researchers have developed various pre-processing techniques like relabelling of ground truth labels termed as massaging [22,23] and modifying other remaining features known as perturbation [15]. These techniques tweak data points to tackle bias, but this essentially creates synthetic data, which makes it hard to ascertain its quality and integrity. Pre-processing techniques also include sampling methods like oversampling [21,37] and undersampling [9,33] which aim to create a balanced distribution in the training data. Oversampling the minority class may cause overfitting while undersampling the majority class can discard useful information. To overcome this, instead of sampling, different weights are assigned to the samples in the reweighing method [5], which makes the model pay more attention to samples from the minority class or samples considered more important.

Further, Chai et al. [8] addresses the issue of uneven sample distribution by proposing an adaptive reweighing method. This method learns adaptive weights for each sample to balance groups across different demographics.

Pre-processing techniques can reduce bias, but they are limited in capturing complex feature interactions. Therefore, researchers have developed in-processing techniques that incorporate fairness into the model architecture, creating models that are naturally fairer and less prone to bias amplification. Regularization approaches like decision trees [24], adding fairness regularisation [26], and meta-algorithm [7] modify the algorithm's objective function to penalize models that discriminate against certain groups. Compositional approaches train multiple classification models, either tailored for specific population groups [6,29] or combined into an ensemble [11,20]. Regularization can reduce bias but often favors the majority class by minimizing overall error. Compositional approaches, which train separate models for each group, avoid this issue but are computationally expensive and impractical for many applications [36]. Researchers also use adversarial learning, where two models are trained together: one predicts the target variable while the other exploits fairness issues in the first. This improves both fairness and accuracy [12]. Researchers have employed adversarial training in various ways. Zhang et al. [39] used logistic regression for both the classifier and adversary. Other methods use a neural network to optimize the prediction of correct labels and minimize the prediction of sensitive information simultaneously [2]. While reweighing is typically a pre-processing step, Petrovic et al. [32] introduced FAIR, an in-processing technique using an adversarial model to learn an instance reweighing function to reduce bias. Beyond adversarial training, representation learning offers another approach that focuses on creating fair representations of the data using techniques like optimization [18], adversarial learning [14], and variational autoencoders (VAEs) [28] as used in F2VAE [4] to promote unbiased recommendations. Representation learning techniques aim to prevent biases by creating new representations that capture relevant task information while suppressing sensitive attributes [38]. Hu et al. [19] introduced FairNN, which promotes fairness by jointly learning a fair data representation using an autoencoder with a KL-divergence constraint that excludes sensitive attributes and a classifier with equalized odds regularization to penalize biased predictions. Our proposed method also learns new representations via autoencoders; however, contrary to FairNN, our method uses gating operations to learn fair latent representations.

While most of the researchers have used any one type of mitigation technique among pre-processing, in-processing, or post-processing, few have also combined one or more types to resolve the challenge of bias mitigation [6,20]. Our work also draws inspiration from these techniques of using mitigation techniques in conjunction. Our proposed method simultaneously leverages a pre-processing technique, adaptive reweighing, and an in-processing technique, fair latent representation learning, to train fair models.

3 Problem Formulation

We have formulated the problem as any binary classification task where we have an input dataset x with n number of samples: $x_1, x_2, ..., x_n$. Each data point x_i is associated with target attribute $y_i \in \{0, 1\}$ and sensitive attribute $s_i \in \{0, 1\}$. The objective is to train a model such that:

- ○ the model shows high fairness, indicating that the output $\hat{y}$ is equitable with respect to the sensitive attribute s_i, and
- ○ the model's accuracy remains intact, meaning the discrepancies between y and $\hat{y}$ are minimal.

Let h_θ be an autoencoder with parameter θ, trained on an input dataset x. h_θ consists of an encoder module that maps the input to a lower dimensional latent representation and a decoder module that uses the latent representation to reconstruct the original data. The latent representation is passed onto two branches with fully connected layers to obtain an updated latent representation (used for the main task prediction) that does not contain the sensitive attribute information. This is done to nullify the effect of the sensitive attribute on final model prediction. While training the model towards fairness goals, it is also important to maintain a decent model performance. The proposed method achieves both objectives using the following.

Fair Latent Representation Learning: The proposed method utilizes gating operations that regulate the flow of information such that the features relevant to the main task are enhanced while the features related to the sensitive attributes are suppressed. Unlike previous methods [10] that drop certain features to address fairness concerns, the proposed method considers all the features and manipulates the importance of each feature for fair outcomes. In essence, the proposed method strikes a balance between accuracy and fairness. By using gating operations to manage feature importance, it ensures that the model learns and emphasizes task-relevant features while actively addressing and mitigating biases associated with sensitive attributes.

Adaptive Reweighing: The adaptive reweighing process involves learning a set of weights for each sample, effectively balancing the representation for different classes within the training data. This weight assignment is achieved through the solution of a convex optimization problem. Inspired by the methodology proposed by Chai and Wang [8], we frame the task as a convex optimization problem. However, our approach differs by considering only the target variable for weight adjustment, not intersectional subgroups. This simplifies the weight learning process while still focusing on achieving a balanced and unbiased representation of the data based on the target variable. As a result, the model becomes adept at recognizing and appropriately considering the importance of each class label, contributing to a more equitable distribution of attention across diverse data samples.

4 Proposed Method

The proposed method aims to learn a fair latent representation while performing adaptive reweighing of input data samples to achieve fair outcomes and improve the overall model performance. Fair latent representation learning focuses on learning a transformation of training data to a latent space, which is fair with respect to sensitive attributes while preserving the maximum information in the training data. On the other hand, the adaptive reweighing process learns a set of weights corresponding to each input sample to balance the representation of different classes. Figure 1 shows the block diagram of model training using the proposed method.

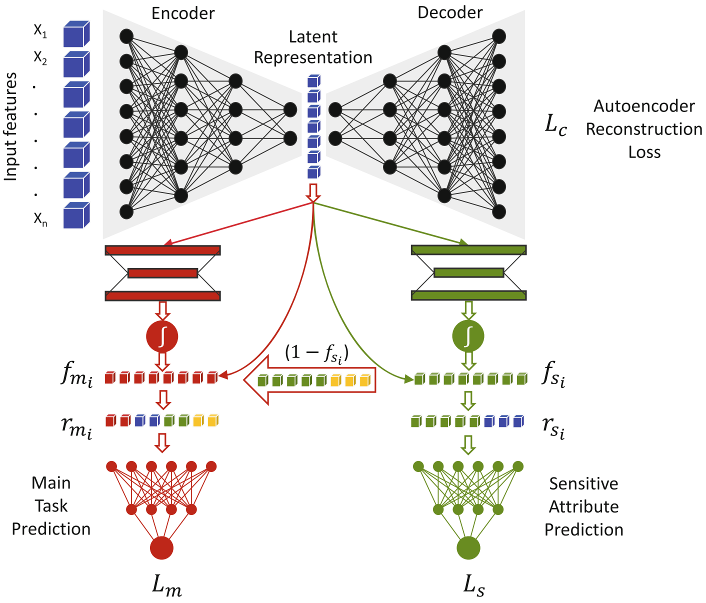


Fig. 1. Block diagram illustrating model training using the proposed method. The autoencoder generates a latent representation from the input features. This latent representation is then fed into two separate branches: Main Task Prediction and Sensitive Attribute Prediction. Each branch consists of a multi-layer perceptron (MLP) with one hidden layer followed by sigmoid activation. The output of the MLP is combined with the latent representation after the encoder module to create new latent representations. These new representations are then fed to the neural networks for the main task and sensitive attribute prediction tasks. The outputs of both the neural networks and the decoder are used to minimize the loss function for model training.

Initially, input data point x_i is given as input to the autoencoder h_θ to obtain the latent representation r_i (after the encoder module) and the reconstructed output $\hat{x_i}$ (after the decoder module). To learn a meaningful latent representation representative of the input data points, reconstruction loss is applied to the decoder output.

$$L_c = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \hat{x}_i)^2} \tag{1}$$

The latent representation r_i obtained after the encoder module of h_θ is passed onto two separate branches for (i) main task prediction and (ii) sensitive attribute prediction. Both the branches consist of a multi-layer perceptron (MLP) with a single hidden layer followed by the sigmoid function. The outputs of both the branches are considered as feature representations for the main task (f_{m_i}) and sensitive attribute prediction task (f_{s_i}). Mathematically, it is represented as:

$$f_{m_i} = Q_m(r_i) \tag{2}$$

$$f_{s_i} = Q_s(r_i) \tag{3}$$

where, Q_m and Q_s represent the MLP for the main task and sensitive attribute prediction task, respectively. The latent representation r_i is combined with the outputs of Q_m and Q_s to obtain the updated representations for each task. The updated representations are then used for main task prediction and sensitive attribute prediction. The latent representations are updated using the following equations:

$$r_{m_i} = f_{m_i} * (1 - f_{s_i}) * r_i \tag{4}$$

$$r_{s_i} = f_{s_i} * r_i \tag{5}$$

where, $*$ represents element-wise multiplication. r_{m_i} and r_{s_i} represent the updated latent representations for the main task and sensitive attribute prediction task, respectively. In Eq. 4, f_{m_i} weigh the features important for the main task, and $(1 - f_{s_i})$ suppress the sensitive attribute features, thereby eliminating the bias-inducing features from the latent representation of the main task. To ensure that the important features related to the sensitive attributes are suppressed during main task prediction, it is crucial that the supervision from the sensitive attribute predictor is meaningful. For this purpose, in Eq. 5, the latent representation of sensitive attribute predictor is updated using f_{s_i} to weigh the features important for sensitive attribute prediction.

The updated latent representations are given as input to two neural networks for the main task prediction and sensitive attribute prediction, respectively. Let L_m be the loss for the main task prediction.

$$L_m = -\left(\sum_{i=1}^{n}\left[y_i \log(p_{m_i}) + (1 - y_i)\log(1 - p_{m_i})\right]\right) \tag{6}$$

where, p_{m_i} is the probability of predicting x_i to one of the classes of the target variable and y_i is the true label. During training the model using L_m, the adaptive reweighing approach is adopted to balance class labels of the target variable by learning a set of weights for each sample x_i. The aim is to focus on the wrongly classified samples to assign more weights to them. The weights are learned by solving the following optimization problem.

$$\max_{w} \sum w_i \; L_m(y_i, \hat{y_i}) - \alpha \sum ||w||_2^2 \tag{7}$$

$$s.t. \sum_{i=0}^{i=n} w_i = c, \;\; w >= 0 \tag{8}$$

where, w_i is the weight assigned to i^{th} sample, $\hat{y_i}$ is the predicted label, α is a hyper-parameter to control the number of samples that obtain non-zero weights, c is a constant. Next, the sensitive attribute predictor is trained using the following loss function.

$$L_s = -\left(\sum_{i=1}^{n} [z_i \log(p_{s_i}) + (1 - z_i) \log(1 - p_{s_i})]\right) \tag{9}$$

where, p_{s_i} is the probability of predicting x_i to one of the classes of the sensitive attribute and z_i is the true label of the sensitive attribute.

The final loss function to train the autoencoder h_θ along with the two branches to learn the updated latent representations followed by predicting the main task and the sensitive attribute is written below.

$$L = \lambda_1 * L_m + \lambda_2 * L_s + \lambda_3 * L_c \tag{10}$$

where, λ_1, λ_2 and λ_3 are the hyper-parameters to weigh different loss terms. The loss L optimizes the model to reduce bias in model prediction and improve the overall model performance.

Fair Model Prediction: The proposed method helps to debias the latent representation for fair model prediction by ensuring that the features relevant to the sensitive attributes are not used during the main task prediction. During testing, the branch for sensitive attribute prediction is removed from the autoencoder

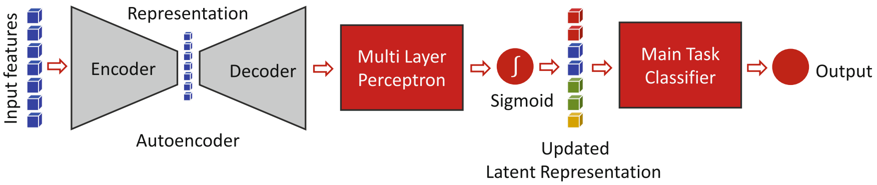


Fig. 2. Illustrating the model architecture during testing. The sensitive attribute prediction branch is removed while making the final prediction.

architecture (Fig. 2). This ensures that only the debiased latent representation, focusing on task-relevant features, is used by the MLP of the main task prediction branch for the final prediction.

5 Experimental Setup

Experiments are performed for classification tasks on three publicly available datasets to evaluate the effectiveness of the proposed method using performance and fairness evaluation metrics. The following discusses the dataset details, evaluation metrics, and implementation details.

Table 1. Details of the Adult, COMPAS, and German datasets.

Dataset	Samples (Train/Test)	Target class	Sensitive attribute	Groups
Adult	32,561/16,281	Income	Sex	Female/Male
COMPAS	5,771/1,443	Two-year recidivism	Race	African-American/Caucasian
German	800/200	Credit score	Sex	Female/Male

5.1 Dataset Details

Experiments are performed on the following publicly available tabular datasets. Table 1 summarizes the dataset details.

Adult Dataset [1]**:** The dataset contains 48,842 samples, with 32,561 samples in the training set and 16,281 in the testing set. It consists of 14 features and one target variable. It is also known as "Census Income" dataset, where the aim is to predict whether income exceeds $50K/yr. We consider *Sex* as the sensitive attribute. The dataset is downloaded from the UCI Machine Learning Repository.

COMPAS Dataset [27]**:** The COMPAS dataset used has 7214 samples and 51 attributes. COMPAS assesses a criminal's likelihood of reoffending for judges and parole officers. We focused on 'Caucasian' and 'African-American' races, resulting in 6150 samples and 30 attributes after preprocessing. Date columns were converted to days, and categorical attributes were one-hot encoded. 'Race', with classes 'African-American' and 'Caucasian', is the sensitive attribute. Data was split randomly into 80-20 sets.

German Credit Risk Dataset [13]**:** The dataset classifies people with good or bad credit risk based on 20 attributes. There are 1000 instances present in the data created by Dr. Hans Hofmann hosted on the UCI Machine Learning Repository. For preprocessing, categorical columns are processed using one-hot encoding, and the train-test split is done using the random 80-20 split. We consider *sex* as the sensitive attribute.

5.2 Evaluation Metrics

We evaluate the effectiveness of the proposed method using both performance and fairness metrics. Performance is evaluated using the ***Accuracy*** of the model for the main task prediction. Additionally, we computed the ***AUC*** and ***F1-scores*** (due to imbalanced training data distribution). For fairness evaluation, we employed the following three metrics:

Disparate Impact (DI) [23] assesses whether a classifier assigns positive classifications at the same rate across different groups. It measures the difference in the positive outcome rate between different groups of a sensitive attribute.

$$DI = |p(\hat{y} = 1 \mid s = 0) - p(\hat{y} = 1 \mid s = 1)| \tag{11}$$

Equal Opportunity (E. Opp) [35] requires that the true positive rate (TPR) be equal across different groups defined by a sensitive attribute s.

$$E.Opp. = |p(\hat{y} = 1 \mid y = 1, s = 0) - p(\hat{y} = 1 \mid y = 1, s = 1)| \tag{12}$$

Equalized Odds (EO) [35] requires that both the true positive rate (TPR) and the false positive rate (FPR) be equal across different groups defined by a sensitive attribute s.

$$\begin{aligned} EO = &|p(\hat{y} = 1 \mid y = 1, s = 0) - p(\hat{y} = 1 \mid y = 1, s = 1)| \\ &+ |p(\hat{y} = 1 \mid y = 0, s = 0) - p(\hat{y} = 1 \mid y = 0, s = 1)| \end{aligned} \tag{13}$$

These definitions represent different aspects of group fairness as applied in real-world scenarios. Equal Opportunity and Equalized Odds relate directly to the model's performance, striving to maintain similar accuracy levels across different sensitive groups. On the other hand, Disparate Impact focuses on achieving balanced outcomes for different groups, such as ensuring that a bank issues loans at equal rates regardless of gender. Lower values of DI, E. Opp., and EO indicate lower bias in model predictions. Ideally, these metrics should be zero for a completely fair model.

5.3 Implementation Details

Experiments are performed using autoencoders of different architectures for different datasets. We employed an autoencoder architecture due to the tabular format of the datasets. However, the proposed algorithm is generalizable across various domains and model architectures.

For the *Adult* dataset, we used a 7-layer autoencoder. The encoder encodes the input data using four dense layers with dimensions of 64, 32, 32, and 16, respectively. The decoder reconstructs the latent representation using three dense layers with dimensions of 32, 32, and 64, respectively. A 5-layer autoencoder is used for the *COMPAS* dataset, where the encoder contains three dense layers of dimension 64, 32, and 16, respectively, while the decoder contains two dense layers of dimension 32 and 64, respectively. For the *German* dataset, we employ a simplified 3-layer autoencoder with dimensions 64, 16, and 64, respectively. All the layers are followed by ReLU activation. The MLP for both the main task prediction and sensitive attribute prediction branches contains three layers with dimensions of 16, 8, and 16, respectively. The output of the MLP is then fed into two separate neural networks for main task prediction and sensitive attribute prediction. Both neural networks contain one hidden layer of dimension 8, followed by the output layer with sigmoid activation function.

The models are trained using the Adam optimizer for 20 epochs for the *Adult* dataset, 100 epochs for the *German* dataset, and 50 epochs for the *COMPAS* dataset. A batch size of 32 is used for all datasets. The learning rate is set to 0.0001 with a decay rate of 0.75 for all datasets. The hyper-parameter c is set to 10000 (experimentally obtained) for all three datasets. Grid search is employed for hyper-parameter (λ_1, λ_2, λ_3) tuning. The values are set to (0.1, 0.8, 10) for the Adult dataset, (0.1, 3, 10) for the COMPAS dataset, and (1, 10, 0.5) for the German dataset corresponding to λ_1, λ_2, λ_3, respectively.

For pre-processing, we used one-hot encoding to convert non-numerical features to numerical features. We further normalize numerical features to zero mean and unit variance. The implementation is carried out using Python 3.8 and popular libraries, including scikit-learn, TensorFlow.

6 Results and Analysis

In this section, we discuss the experimental results to demonstrate the effectiveness of the proposed method towards bias mitigation. We first compare the proposed method with a few baseline methods to illustrate that model training using the proposed method leads to fair model predictions while maintaining decent accuracy, unlike conventional machine learning algorithms that optimize the model with the objective of maximizing the overall performance, ignoring fairness constraints. The baseline methods include (i) logistic regression (LR), (ii) multi-layer perception (MLP), and (iii) autoencoder with classifier (AE) models. The selection of LR and MLP is made to encompass a comprehensive spectrum of complexity within different baseline methods. AE is included in baseline methods because the proposed method's architecture is also based on the autoencoder model. Tables 2, 3, and 4 summarize results corresponding to the Adult, COMPAS, and German datasets, respectively. It is observed that baseline methods perform well in terms of accuracy, but in most cases, they do not perform well on fairness metrics. For instance, the accuracy of AE on the Adult dataset is 85.58%. However, the DI, E. Opp, and EO of the model are 0.17, 0.07, and 0.14, respectively. The high values of these metrics indicate that the model prediction is biased across different groups.

Since baseline methods fail to train fair models, we compare the proposed method with existing approaches that explicitly mitigate bias in model prediction, including fairness with adaptive weights (FAW) [8], FAIR [32], and FairNN [19]. These approaches are chosen for comparison because they are similar to

Table 2. Experimental results on the Adult dataset using existing and the proposed method. Accuracy is reported in (%).

	LR	MLP	AE	FAIR	FairNN	FAW	Ours
Accuracy↑	85.41	85.47	85.58	82.89	85.36	82.44	85.73
AUC↑	0.90	0.91	0.91	0.86	0.91	0.91	0.91
F1 Score↑	0.66	0.67	0.67	0.59	0.64	0.63	0.64
DI↓	0.17	0.17	0.17	0.13	0.15	0.17	0.12
E. Opp.↓	0.07	0.06	0.07	0.02	0.07	0.02	0.01
EO↓	0.14	0.13	0.14	0.07	0.13	0.08	0.05

Table 3. Experimental results on the COMPAS dataset using existing and the proposed method. Accuracy is reported in (%).

	LR	MLP	AE	FAIR	FairNN	FAW	Ours
Accuracy↑	96.21	95.50	94.80	93.58	96.18	97.89	96.50
AUC↑	0.98	0.97	0.97	0.96	0.97	1.00	0.97
F1 Score↑	0.96	0.95	0.95	0.93	0.96	0.98	0.96
DI↓	0.10	0.09	0.09	0.09	0.10	0.10	0.09
E. Opp.↓	0.005	0.02	0.01	0.004	0.02	0.02	0.001
EO↓	0.02	0.04	0.03	0.01	0.02	0.02	0.01

our proposed method in terms of mitigating bias by learning fair representations and/or reweighing the data instances. FAW aims to achieve group-level balance among different demographic groups by learning adaptive weights for each sample. In contrast, our method learns adaptive weights to balance target class labels. FAIR utilizes adversarial training to learn a reweighing function for training data instances to reduce the impact of biased instances. FairNN jointly trains an autoencoder with KL-divergence constraint and a classifier with equalized odds regularization to learn fair representation. On the other hand, our method uses a gating mechanism to learn fair representation and adaptive reweighing to improve overall model performance. While the original FAIR and FairNN results are reported for the Adult dataset and a few others, they are not available for the COMPAS and German datasets. To ensure an equitable comparison, we have extended the methodologies of FAIR and FairNN to encompass the COMPAS and German datasets as well. Additionally, FAW used a different version of the COMPAS dataset compared to the standard one available on the ProPublica website [30]. Hence, we executed their publicly accessible code on the standard COMPAS datasets to ensure a fair comparison.

Our method outperforms existing approaches on the Adult dataset, achieving the lowest values for all three fairness metrics and higher accuracy. Similar performance is observed on the COMPAS and German datasets as well. This indicates that the model trained using the proposed method is able to predict fair outcomes without compromising model performance. The loss function for

Table 4. Experimental results on the German Credit Risk dataset using existing and the proposed method. Accuracy is reported in (%).

	LR	MLP	AE	FAIR	FairNN	FAW	Ours
Accuracy↑	73.50	72.50	75.00	73.50	74.00	74.50	76.50
AUC↑	0.85	0.77	0.81	0.81	0.81	0.79	0.78
F1 Score↑	0.79	0.82	0.84	0.80	0.83	0.80	0.84
DI↓	0.06	0.06	0.04	0.06	0.04	0.03	0.05
E. Opp.↓	0.02	0.04	0.01	0.03	0.04	0.09	0.002
EO↓	0.06	0.09	0.06	0.05	0.06	0.19	0.01

the proposed model has been designed in such a way that the accuracy is not affected at the cost of the fairness. The loss function has a separate term for accuracy ensuring that it is also optimised along with the fairness terms. For further analysis, we have compared the latent representation after the encoder module of the autoencoder with the updated latent representation obtained after eliminating bias-inducing features for main task prediction using PCA visualizations [31]. Figure 3 shows the PCA visualizations of both the latent representations. It is observed that the latent representation after the encoder module is clearly separable by gender. However, after suppressing the sensitive attribute-related features, the updated latent representation is no longer separable by gender, showcasing unbiased latent representation for main task prediction.

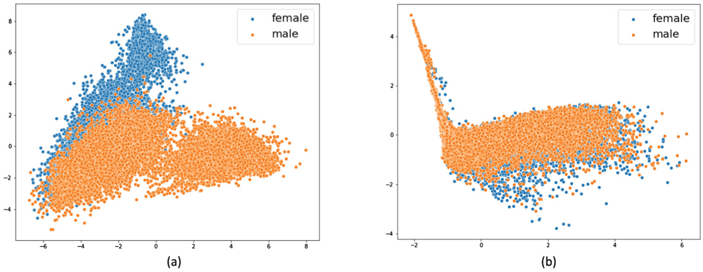


Fig. 3. PCA visualization of the (a) latent representation after the encoder layer and (b) updated latent representation on the Adult dataset. The visualization in (a) shows that the female and male groups are separable, which becomes inseparable in (b) after suppressing features related to the sensitive attribute.

7 Ablation Study

We have conducted an experiment by disabling the sensitive attribute prediction branch to assess its impact on learning a fair latent representation for the main task. In other words, the model is trained without supervision from the sensitive attribute prediction branch. The latent representation from the encoder layer is combined with the output of the main task prediction branch's MLP to update it. This updated latent representation is used by the main task prediction classifier.

Table 5 shows the results of the ablated model on the Adult dataset. Comparison is performed with the proposed method. It is observed that the ablated model performs poorly on fairness metrics. For instance, DI shows an approximate 25% increase, whereas E. Opp. and EO rise to 0.06 and 0.15, respectively, in comparison to the proposed method. This highlights the importance of the sensitive attribute prediction branch in providing information related to bias-inducing features that need to be suppressed during main task prediction.

Table 5. Results on the Adult dataset after ablating sensitive attribute prediction branch. Comparison is made with the proposed method. Accuracy is reported in (%).

	Accuracy↑	AUC↑	F1 Score↑	DI↓	E. Opp.↓	EO↓
Without Sensitive Attribute Branch	85.34	0.75	0.64	0.15	0.06	0.15
Ours	85.73	0.91	0.64	0.12	0.01	0.05

8 Conclusion

As AI decision-making systems become more pervasive, concerns regarding potential bias and unfair outcomes have gained significant attention. This research delves into two primary sources of such bias: imbalanced training data and correlations between sensitive attributes and the target variable. To this end, we propose a novel method to train fair models. The proposed method employs fair representation learning to guide the model towards features that are genuinely relevant to the task at hand, effectively reducing the influence of features that might be implicitly linked to sensitive attributes and potentially leading to biased outcomes. Furthermore, we address the issue of data imbalance within the training data through adaptive reweighing, ensuring that all target class labels are balanced during model training.

Evaluations conducted on benchmark datasets demonstrate the effectiveness of the proposed method in bias mitigation. The ablation study reinforced the significance of the supervision provided by the sensitive attribute branch; removing this branch during model training resulted in a drop in fairness metrics. In the current experimental setup, the model is optimized for a single downstream task. As part of future work, the proposed method can be extended to debias latent representations for multiple downstream tasks to mitigate bias due to multiple sensitive attributes.

References

1. Becker, B., Kohavi, R.: Adult. UCI Machine Learning Repository (1996). https://doi.org/10.24432/C5XW20
2. Beutel, A., Chen, J., Zhao, Z., Chi, E.H.: Data decisions and theoretical implications when adversarially learning fair representations (2017)
3. Bolukbasi, T., Chang, K.W., Zou, J.Y., Saligrama, V., Kalai, A.T.: Man is to computer programmer as woman is to homemaker? Debiasing word embeddings. In: Advances in Neural Information Processing Systems, vol. 29 (2016)
4. Borges, R., Stefanidis, K.: F2VAE: a framework for mitigating user unfairness in recommendation systems. In: 37th Association for Computing Machinery/Special Interest Group on Applied Computing Symposium on Applied Computing, pp. 1391–1398 (2022)
5. Calders, T., Kamiran, F., Pechenizkiy, M.: Building classifiers with independency constraints. In: International Conference on Data Mining Workshops, pp. 13–18. IEEE (2009)
6. Calders, T., Verwer, S.: Three Naive Bayes approaches for discrimination-free classification. Data Mining Knowl. Discov. **21**, 277–292 (2010)
7. Celis, L.E., Huang, L., Keswani, V., Vishnoi, N.K.: Classification with fairness constraints: a meta-algorithm with provable guarantees. In: Conference on Fairness, Accountability, and Transparency, pp. 319–328 (2019)
8. Chai, J., Wang, X.: Fairness with adaptive weights. In: International Conference on Machine Learning, pp. 2853–2866. Proceedings of Machine Learning Research (2022)
9. Chakraborty, J., Majumder, S., Yu, Z., Menzies, T.: Fairway: a way to build fair ml software. In: ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering, pp. 654–665 (2020)
10. Chaudhary, B., Pandey, A., Bhatt, D., Tiwari, D.: Practical bias mitigation through proxy sensitive attribute label generation. arXiv preprint arXiv:2312.15994 (2023)
11. Chen, Z., Zhang, J.M., Sarro, F., Harman, M.: Maat: a novel ensemble approach to addressing fairness and performance bugs for machine learning software. In: Joint European Software Engineering Conference and Symposium on the Foundations of Software Engineering. Association for Computation Machinery (2022)
12. Dalvi, N., Domingos, P., Sanghai, S., Verma, D.: Adversarial classification. In: ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 99–108. Association for Computation Machinery (2004)
13. Dua, D., Graff, C., et al.: UCI machine learning repository, vol. 7, no. 1, p. 62 (2017). http://archive.ics.uci.edu/ml
14. Edwards, H., Storkey, A.: Censoring representations with an adversary (2015)
15. Feldman, M., Friedler, S.A., Moeller, J., Scheidegger, C., Venkatasubramanian, S.: Certifying and removing disparate impact. In: ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 259–268. Association for Computer Machinery (2015)
16. Friedler, S.A., Scheidegger, C., Venkatasubramanian, S., Choudhary, S., Hamilton, E.P., Roth, D.: A comparative study of fairness-enhancing interventions in machine learning. In: Conference on Fairness, Accountability, and Transparency, pp. 329–338 (2019)
17. Gupta, A., Bhatt, D., Pandey, A.: Transitioning from real to synthetic data: quantifying the bias in model (2021)
18. Hacker, P., Wiedemann, E.: A continuous framework for fairness (2017)

19. Hu, T., Iosifidis, V., Liao, W., Zhang, H., Yang, M.Y., Ntoutsi, E., Rosenhahn, B.: Fairnn-conjoint learning of fair representations for fair decisions. In: Discovery Science: 23rd International Conference, Thessaloniki, Greece, pp. 581–595. Springer, Cham (2020)
20. Iosifidis, V., Fetahu, B., Ntoutsi, E.: Fae: a fairness-aware ensemble framework. In: International Conference on Big Data, pp. 1375–1380. Institute of Electrical and Electronics Engineers (2019)
21. Iosifidis, V., Ntoutsi, E.: Dealing with bias via data augmentation in supervised learning scenarios. Jo Bates Paul D. Clough Robert Jäschke **24**(11) (2018)
22. Kamiran, F., Calders, T.: Classifying without discriminating. In: International Conference on Computer, Control and Communication, pp. 1–6. IEEE (2009)
23. Kamiran, F., Calders, T.: Data preprocessing techniques for classification without discrimination. Knowl. Inf. Syst. **33**(1), 1–33 (2012)
24. Kamiran, F., Calders, T., Pechenizkiy, M.: Discrimination aware decision tree learning. In: International Conference on Data Mining, pp. 869–874. Institute of Electrical and Electronics Engineers (2010)
25. Kamiran, F., Mansha, S., Karim, A., Zhang, X.: Exploiting reject option in classification for social discrimination control. Inf. Sci. **425**, 18–33 (2018)
26. Kamishima, T., Akaho, S., Sakuma, J.: Fairness-aware learning through regularization approach. In: International Conference on Data Mining Workshops, pp. 643–650. Institute of Electrical and Electronics Engineers (2011)
27. Larson, J., Angwin, J., Kirchner, L., Mattu, S.: How we analyzed the compas recidivism algorithm. ProPublica (2016). https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm
28. Louizos, C., Swersky, K., Li, Y., Welling, M., Zemel, R.: The variational fair autoencoder. In: Bengio, Y., LeCun, Y. (eds.) International Conference on Learning Representations (2016)
29. Oneto, L., Doninini, M., Elders, A., Pontil, M.: Taking advantage of multitask learning for fair classification. In: AAAI/ACM Conference on AI, Ethics, and Society, pp. 227–237 (2019)
30. Pagano, T.P., et al.: Bias and unfairness in machine learning models: a systematic literature review (2022)
31. Pearson, K.: Liii. on lines and planes of closest fit to systems of points in space. London Edinburgh Dublin Philos. Mag. J. Sci. **2**(11), 559–572 (1901)
32. Petrović, A., Nikolić, M., Radovanović, S., Delibašić, B., Jovanović, M.: Fair: fair adversarial instance re-weighting. Neurocomputing **476**, 14–37 (2022)
33. Salimi, B., Rodriguez, L., Howe, B., Suciu, D.: Interventional fairness: causal database repair for algorithmic fairness. In: International Conference on Management of Data, pp. 793–810. Association for Computer Machinery (2019)
34. Sharma, S., Zhang, Y., Ríos Aliaga, J.M., Bouneffouf, D., Muthusamy, V., Varshney, K.R.: Data augmentation for discrimination prevention and bias disambiguation. In: AAAI/ACM Conference on AI, Ethics, and Society, pp. 358–364 (2020)
35. Wan, M., Zha, D., Liu, N., Zou, N.: In-processing modeling techniques for machine learning fairness: a survey. ACM Trans. Knowl. Discov. Data **17**(3), 1–27 (2023)
36. Yang, J., Soltan, A.A., Eyre, D.W., Yang, Y., Clifton, D.A.: An adversarial training framework for mitigating algorithmic biases in clinical machine learning. NPJ Digit. Med. **6**(1), 55 (2023)
37. Zelaya, V., Missier, P., Prangle, D.: Parametrised data sampling for fairness optimisation. KDD XAI (2019)

38. Zemel, R., Wu, Y., Swersky, K., Pitassi, T., Dwork, C.: Learning fair representations. In: Dasgupta, S., McAllester, D. (eds.) International Conference on Machine Learning, vol. 28, pp. 325–333. Proceedings of Machine Learning Research (2013)
39. Zhang, B., Lemoine, B., Mitchell, M.: Mitigating unwanted biases with adversarial learning. In: AAAI/ACM Conference on AI, Ethics, and Society, pp. 335–340. Association for Computation Machinery (2018)

FineFACE: Fair Facial Attribute Classification Leveraging Fine-Grained Features

Ayesha Manzoor and Ajita Rattani(✉)

Department of Computer Science and Engineering, University of North Texas at Denton, Denton, TX, USA
ayeshamanzoor@my.unt.edu, ajita.rattani@unt.edu

Abstract. Published research highlights the presence of demographic bias in automated facial attribute classification algorithms, particularly impacting women and individuals with darker skin tones. Existing bias mitigation techniques typically require demographic annotations and often obtain a trade-off between fairness and accuracy, i.e., Pareto inefficiency. Facial attributes, whether common ones like gender or others such as "chubby" or "high cheekbones", exhibit high interclass similarity and intraclass variation across demographics leading to unequal accuracy. This requires the use of local and subtle cues using fine-grained analysis for differentiation. This paper proposes a novel approach to fair facial attribute classification by framing it as a fine-grained classification problem. Our approach effectively integrates both low-level local features (like edges and color) and high-level semantic features (like shapes and structures) through cross-layer mutual attention learning. Here, shallow to deep CNN layers function as experts, offering category predictions and attention regions. An exhaustive evaluation on facial attribute annotated datasets demonstrates that our FineFACE model improves accuracy by 1.32% to 1.74% and fairness by 67% to 83.6%, over the SOTA bias mitigation techniques. Importantly, our approach obtains a Pareto-efficient balance between accuracy and fairness between demographic groups. In addition, our approach does not require demographic annotations and is applicable to diverse downstream classification tasks. To facilitate reproducibility, the code and dataset information is available at https://github.com/VCBSL-Fairness/FineFACE.

Keywords: Fairness in AI · Facial Attribute Classification · Fine-grained Features

Supported by National Science Foundation, USA.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/978-3-031-78110-0_6.

A. Antonacopoulos et al. (Eds.): ICPR 2024, LNCS 15329, pp. 81–97, 2025.
https://doi.org/10.1007/978-3-031-78110-0_6

1 Introduction

Automated facial analysis algorithms encompass face detection, face recognition, and facial attribute classification (such as gender, race, high cheekbones, and attractiveness) [8,17,31]. These algorithms are deeply integrated into various sectors, such as surveillance and border control, retail and entertainment, healthcare, and education.

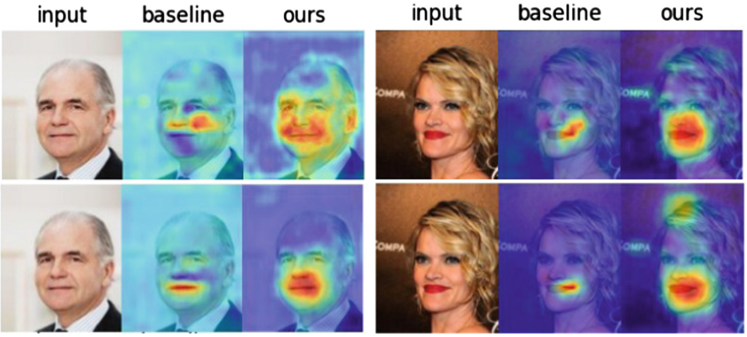


Fig. 1. Visualization of the attention map obtained by our proposed FineFACE over baseline (both using ResNet50 backbone) for facial attribute classification. The highly activated region is shown by the red zone on the map, followed by yellow, green, and blue zones. Top: "High Cheekbones" classifier. Bottom: "Smiling" classifier. (Color figure online)

Numerous existing studies [1,12,26] investigating the *fairness of facial attribute classification* algorithms confirm the presence of *performance disparities between demographic groups, such as gender and race.* Thus, bias in these algorithms emerges as a significant societal issue that warrants immediate redress, particularly for the large-scale deployment of fair and trustworthy systems across demographics. In this direction, the vision community has proposed several bias mitigation techniques to address the performance disparities of facial attribute classifiers. Established bias mitigation techniques utilize regularization [13], attention mechanism [21], adversarial debiasing [3,33], GAN-based over-sampling [25,36], multi-task classification [5], and network pruning [15].

These existing bias mitigation techniques often require demographically annotated training sets and are limited in their generalizability. Importantly, these techniques often sacrifice overall classification accuracy in pursuit of improved fairness, making them *Pareto inefficient* [33,36]. It was demonstrated in [36] that fairness violations in vision models are largely driven by the variance component of bias-variance decomposition. Consequently, *one effective way to improve fairness is by decreasing the variance within each demographic subgroup by focusing on local and subtle cues.* This can be obtained through learning enhanced **feature representation** for each demographic subgroup, also supported by [4,12].

Following this line of thought, enhancing feature representation for each demographic subgroup is crucial in improving fairness without compromising overall performance. Traditional facial attribute classifiers [2,3,12,23] rely predominantly on high-level discriminative and semantically meaningful information often obtained from the final layers of the deep convolutional neural network (CNN). However, the lower layers of the deep learning model capture (low-level) essential features and patterns in faces vital for attribute classification, such as **(a)** facial contours and edges, including the outline of the face, jawline and cheekbones, **(b)** texture of facial regions, such as skin and hair, **(c)** position and shape information, and **(d)** lighting condition and its effect on the appearance of facial features. Integrating low-level details from the lower layers of the model will capture local and detailed cues in the learned feature representation.

In our quest to identify these subtle and local cues for learning enhanced feature representation, we **aim** to leverage fine-grained analysis, integrating both high- and low-level features, toward fair facial attribute classification, FineFACE. This is facilitated through a cross-layer mutual attention learning technique that learns fine-grained features by considering the layers of a deep learning model from shallow to deep as independent 'experts' knowledgeable about low-level detailed to high-level semantic information, respectively. These experts are trained in leveraging mutual data augmentation to incorporate attention regions proposed by other experts. An ordinary deep learning model can be considered to use only the deepest expert (using high-level semantic information) for classification. In contrast, our method consolidates the prediction of the categorical label and the attention region of each expert for the final facial attribute classification task.

Figure 1 shows the final CAM visualization obtained by our proposed FineFACE model based on the ResNet50 backbone, for facial attribute classification with "high cheekbone" and "smiling" as target variable using the CelebA dataset [17]. The highly activated region is shown by the red zone on the map, followed by yellow, green, and blue zones in the attention map. For cross-comparison, the visualization of the baseline ResNet50 is also shown for the same classification task. As illustrated in the maps, our FineFace model captures additional information, such as the contours of facial regions derived from the lower layers of the model, leading to enhanced feature representation and, hence, accurate and fair facial attribute classification.

Contributions. In summary, the contributions of our work are as follows: **(i)** We approach fair facial attribute classification from a novel perspective by reformulating it as a fine-grained classification task, **(ii)** We propose a novel approach based on cross-layer mutual attention learning where the prediction is consolidated from shallow (using low-level details) to deep layers (using high-level semantic details) regarded as an independent experts, **(iii)** Extensive evaluation on facial attribute annotated datasets namely, FairFace [11], UTKFace [34], LFWA+ [17], and CelebA [18], and **(iv)** Cross-comparison with the existing bias mitigation techniques, demonstrating the efficacy of our approach in terms of significant improvement in fairness as well as classification accuracy. Thus, obtaining state-of-the-art Pareto-efficient performance.

2 Related Work

In this section, we review the related academic literature.

Bias Mitigation of Facial Attribute Classification. Many studies have highlighted the systematic limitations of facial attribute classification (such as gender, race, and age) between gender-racial groups [2,12,23]. Studies in [3,27] reported bias of facial attribute classification for attractive, smiling, and wavy hair as the target attributes across genders. Following this study, [36] reported bias of gender-independent target attributes, such as black hair, smiling, slightly open mouth, and eyeglasses, between genders. Consequently, numerous strategies have been proposed to mitigate bias [22]. [5] explored the joint classification of gender, age, and race by proposing a multi-task network. [33] included a variable for the group of interest and simultaneously learned a predictor and an adversary via adversarial debiasing. [13] leveraged the power of semantic preserving augmentations at the image level in a self-consistency setting for fair gender classification tasks. [24] introduced a framework that integrates fairness constraints directly into the loss function using Lagrangian multipliers for fair classification. [3] proposed "fair mixup," a data augmentation technique by interpolating data points that improve the generalization of the classifiers trained under group fairness constraints. [25] adopted structured learning techniques using deep-views of the training samples generated using GAN-based latent code editing to improve the fairness of the gender classifier. GAN-based SMOTE "g-SMOTE" was proposed by [36] to strategically enhance the training set for underrepresented subgroups to mitigate bias.

Fine-Grained Visual Classification. Fine-grained classification is a challenging research task in computer vision, which captures the local discriminative features using attention learning [6,35] to distinguish different fine-grained categories. In addition to methods based on attention mechanisms, second-order pooling methods utilize the second-order statistics of deep features to compose powerful representations such as combined feature maps [14] and covariance among deep features [28] for fine-grained classification. Studies have also been proposed to use features or information learned from different layers within a CNN backbone for fine-grained classification. [32] proposed a multi-layered Deconvnet for gaining insight into the functions of intermediate feature layers. They discovered that shallow layers capture low-level details, whereas deep layers capture high-level information. [10] proposed the LayerCAM, which indicates the discriminative regions used by the different layers of a CNN to predict a specific category. Inspired by these two works, [16] proposed the CMAL-Net, which focuses on using attention regions predicted by layers of different depths to mark the cues they learned, letting layers of varying depths to learn from each other's knowledge to improve overall performance.

3 Proposed Methodology

In this section, we elaborate on our proposed FineFACE model based on learning features from different layers of the CNN using the attention mechanism and mutual learning, following the foundational works in [10,16,32].

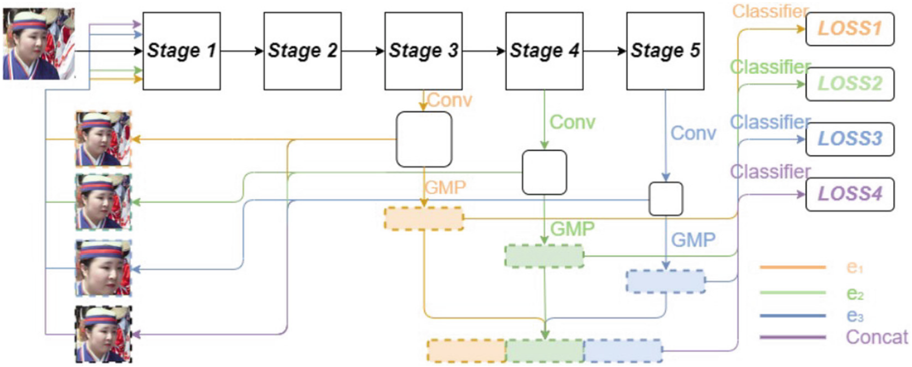


Fig. 2. FineFACE network structure. This figure illustrates this method by introducing three experts e_1, e_2, e_3, on a 5-stage backbone CNN (e.g., ResNet50). The working of each expert and the concatenation of experts are depicted in different colors. Each expert receives feature maps from a specific layer as input and generates a categorical prediction along with an attention region, which is used for data augmentation by other experts. This architecture is trained in multiple steps within each iteration. We start by training the deepest expert (e3), followed by the shallower experts. Finally, in the last step, we train the concatenation of experts to enhance overall performance.

3.1 Expert Construction: Using Shallow to Deep Layers

In this subsection, the construction of experts from shallow to deep layers. Any state-of-the-art CNNs, such as ResNet50, Res2NeXt50, etc. can be used as the backbone CNN denoted by β. β has M layers, and $\{l_1, l_2, ..., l_m, ..., l_M\}$ denote the layers of β from shallow to deep (except the fully connected layers). $\{e_1, e_2, ..., e_n, ..., e_N\}$ are N experts based on these M layers. Each expert encompasses layers from the first layer up to a certain layer such that - e_n consists of the layers from l_1 to l_{m_n}, and $1 \leq m_n \leq$ M. The experts $\{e_1, e_2, ..., e_n, ..., e_N\}$ progressively cover deeper layers of the backbone CNN, and e_N, the deepest expert, covers all layers from l_1 to l_M.

Let $\{x_1, x_2, ..., x_n, ..., x_N\}$ denote the intermediate feature maps produced by β for the experts $\{e_1, e_2, ..., e_n, ..., e_N\}$, respectively. $x_n \in R^{H_n \times W_n \times C_n}$ and H_n, W_n and C_n denote the height, width, and number of channels, respectively. A set of functions $\{F_1(.), F_2(.), ..., F_n(.), ..., F_N(.)\}$ are used to respectively compress $\{x_1, x_2,..., x_n, ..., x_N\}$ into 1D vectorial descriptors $\{v_1, v_2, ..., v_n, ..., v_N\}$, and $v_n \in R^{C_v}$. C_v denotes the length of the 1D vectorial descriptors, and these descriptors given by various experts are of the same length. The $F_n(.)$ for processing x_n is defined as:

$$v_n = F_n(x_n) = f^{GMP}(x_n^{''}), \tag{1}$$

$$x_n^{''} = f^{Elu}(f^{bn}(f^{conv}_{3\times 3\times C_{v/2}\times C_v}(x_n^{'}))), \tag{2}$$

$$x_n^{'} = f^{Elu}(f^{bn}(f^{conv}_{1\times 1\times C_n\times C_{v/2}}(x_n))), \tag{3}$$

where $f^{GMP}(.)$ denotes the Global Max Pooling. $f^{conv}(.)$ depicts the 2D convolution operation by its kernel size. $f^{bn}(.)$ and $f^{Elu}(.)$ denote batch normalization and Elu operations respectively. $x_n^{'}$ and $x_n^{''}$ are intermediate feature maps produced by e_n. Thereafter, $x_n^{''} \in R^{H_n\times W_n\times C_n}$ is used to generate the attention region of e_n as described in Subsect. 3.2. $\{p_1,\, p_2,\, ...,\, p_n,\, ...,\, p_N\}$ denote the prediction scores given by different experts, obtained as $p_n = f_n^{clf}(v_n)$, where $f_n^{clf}(.)$ denotes a fully connected layer-based classifier.

Apart from the prediction scores obtained by the various experts, an overall prediction score is also generated by combining the information from different experts. Specifically, $\{v_1,\, v_2,\, ...,\, v_n,\, ...,\, v_N\}$ are first concatenated for an overall descriptor v_{oval} as: $v_{oval} = f_{concat}(v_1,\, v_2,\, ...,\, v_n,\, ...,\, v_N)$, where $f_{concat}(.)$ denotes concatenation operation. Then v_{oval} is processed into an overall prediction score p_{oval} by a fully connected layer-based classifier as $p_{oval} = f_{oval}^{clf}(v_{oval})$

3.2 Attention Region Prediction

As mentioned above, $x_n^{''}$ denotes an intermediate feature map generated by the expert e_n. We move with an assumption that the classification problem is K-class, and $k_n \in$ 1, 2, ..., K is the category predicted by expert e_n and $x_n^{''} \in R^{H_n\times W_n\times C_n}$. The generation of the attention region proposed by e_n is initialized by producing the class activation map (CAM), which specifies the discriminative image region, for the category k_n based on $x_n^{''}$ specified. The CAM Ω_n ($\Omega_n \in R^{H_n\times W_n}$) produced by the expert e_n is defined as: $\Omega_n(\alpha,\beta) = \sum\limits_{c=1}^{c_v} p_n^c x_n^{''c}(\alpha,\beta)$, where the coordinates (α, β) denote the spatial location of $x_n^{''}$ and $\Omega_n.p_n$ denotes the parameters of $f_n^{clf}(.)$ corresponding to the predicted category k_n. Then, after obtaining Ω, an attention map $\tilde{\Omega}_n \in R^{H_{in}\times W_{in}}$ (H_{in}, W_{in} are the height and width of the input image, respectively) is generated by upsampling Ω_n using a bilinear sampling kernel. Thereafter, $\tilde{\Omega}_n$ is applied with min-max normalization, and each spatial element of the normalized attention map $\tilde{\Omega}_n^{norm}$ is obtained by

$$\tilde{\Omega}_n^{norm}(\alpha,\beta) = \tilde{\Omega}_n(\alpha,\beta) - min(\tilde{\Omega}_n)/max(\tilde{\Omega}_n) - min(\tilde{\Omega}_n). \tag{4}$$

The regions that the expert e_n considers discriminative can be found and cropped by generating a mask $\tilde{\Omega}_n^{mask}$ by setting the elements in $\tilde{\Omega}_n^{norm}$ to 1 for values greater than a threshold t ($t \in [0,\, 1]$) and 0 for the others.

Then, a box that covers all the positive regions of $\tilde{\Omega}_n^{mask}$ is located and cropped from the input image. The cropped region is upsampled to the input image's size and the upsampled attention region A_n is considered as the attention region predicted by e_n and also as data augmentation for remaining experts. Apart from the attention regions proposed by various experts, an overall attention region A_{oval} is generated by summing up the attention information learned by different experts.

3.3 Multi-step Mutual Learning

The experts are trained using progressive multi-step strategy with cross-entropy loss. In the early steps, these experts are trained one by one, which allows them to "focus on" learning the clues of their own expertise without being diverted by other experts. In the last two steps, the experts get together to learn impactful information from the attention regions and the raw image, respectively. Specifically, every iteration of the training takes place in $N + 2$ steps, and in the first N steps, each expert is gradually trained from deep to shallow. In the first step, the deepest expert e_N is trained. Since the training of N involves the experts shallower than e_N, the attention regions proposed by all the experts and the overall attention region $\{A_1, A_2, ..., A_n, ..., A_N, A_{oval}\}$ are also generated at this step. These attention regions showcase the"specialized knowledge" of the experts by highlighting the basis on which each expert made its classification.

From step 2 to N, there is a progression to shallower experts by randomly selecting one input from a pool of images comprising of the raw input and the attention regions proposed by the other experts. The shallow experts rely on the attention regions proposed by deeper experts to learn semantic visual clues (e.g., eyes, nose, and mouth), while the deep experts take the help of shallow experts by learning low-level visual cues (e.g., facial contours like jawline, cheekbones, etc.) from their proposed attention regions. In step $N + 1$, all the experts and their concatenation are trained with the overall attention region A_{oval} in one pass. This step enforces all experts to work together and study the attention information they have combinedly gained for learning more fine-grained features. At step $N + 2$, the concatenation of all the experts is trained with the raw input to make sure the parameters of $f_{oval}^{clf}(.)$ fit the resolution of the original input. The **algorithm** for the multi-step mutual learning strategy is included in Section 2 of the supplementary material.

Inference Phase: Figure 2 illustrates the inference stage of the proposed FineFACE model with $N + 1$ classifiers. For an input image during the inference, $N + 1$ prediction scores are produced by the proposed architecture. For each test image, the raw input and overall attention region are successively fed to the model obtaining $2 \times (N + 1)$ number of prediction scores. The final prediction score for the inference is the average of the $2 \times (N + 1)$ scores. This inference strategy maximizes the classification accuracy as well as fairness of the trained model due to obtaining two kinds of complementary information: (a) information from the prediction scores from various experts and the overall prediction score, and (b) the information from the raw input and overall attention region.

4 Experimental Details

We conducted two sets of **experiments (1)** a face-based gender classifier with gender as the target attribute and race and gender as the protected attributes following studies in [13,25]. **(2)** 13 gender-independent facial attribute classifiers following studies in [3,27,36] with "bags under eyes", "bangs", "black hair", "blond hair", "brown hair", "chubby", "eyeglasses", "gray hair", "high cheekbones", "mouth slightly open", "narrow eyes", "smiling", and "wearing hat" as the 13 gender independent target attributes and gender as the protected attribute. We used the mean scores of these 13 attribute classifiers, following studies in [3,27,36].

4.1 Datasets and Training Protocol

We used standard benchmark datasets widely adopted for evaluating fairness of facial attribute classifiers [13,27,36], namely, FairFace [11], UTKFace [34], LFWA+ [17], and CelebA [18]. In line with existing studies [13,25], a face-based gender classifier was trained on FairFace and evaluated on FairFace, UTKFace, LFWA+, and CelebA (40 attributes). Unlike UTKFace and LFWA+, CelebA does not have race annotations. Hence, we used only the gender attribute for CelebA. For the 13 gender-independent facial attribute classifiers, we used the CelebA (40 attributes) dataset for training and validation. For the fair comparison with the existing studies on fairness [3,27,36], we used only 13 gender-independent attributes from CelebA. Note that protected attribute annotation information is not used during the model training stage, but solely for the purpose of fairness evaluation of the facial attribute classifiers. Additional details on these datasets are given in Table 1.

Table 1. Dataset details including the number of images and demographic groups

Dataset	Images	Demographic groups
FairFace	108K	White, Black, Indian, Asian, Southeast Asian, Middle Eastern, Latino Hispanic
UTKFace	20K	White, Black, Indian, Asian, Others
LFWA+	13K	White, Black, Indian, Asian, Undefined
CelebA	202K	Not Available (only gender information available)

4.2 Implementations Details

For a fair comparison with studies in [12,25,36], we utilized ResNet50 [9] as our method's backbone CNN architecture. The layers of ResNet50, excluding the fully connected layers, are grouped into 5 stages where each stage is a group of layers operating on feature maps of the same spatial size. We use these stages as building blocks for our experts: e_1 encompasses layers from stage 1 to stage 3, e_2 encompasses layers from stage 1 to stage 4, and e_3 encompasses layers from stage 1 to stage 5. In general, the number of stages in a model can be determined by grouping layers operating on the feature maps of the same spatial size, and accordingly, experts can be formed.

We trained all the models used in this study using Stochastic Gradient Descent (SGD) with the number of epochs determined using an early stopping mechanism, the momentum of 0.9, weight decay of 5×10^{-4}, and a mini-batch size of 16 determined using empirical evidence. The learning rate was set as 0.002 with cosine annealing [19]. We fixed the input image size as 448×448, following the common settings in existing fairness studies [7,20]. The threshold t, which is used to generate a mask for the attention region, was set to 0.5 (see Sect. 3.2). We also conducted an *ablation study* of the related design choices (a) different pooling methods for building experts, and (b) the contribution of fusing the prediction scores. See **ablation studies** in Section 3 of the supplementary material for more details.

4.3 Evaluation Metrics

For the gender classifier, the standard evaluation metrics, namely, classification accuracy, Max-Min ratio (the ratio of the best and worst performing subgroups), and Degree of Bias (DoB) (standard deviation of accuracy) are used for fair comparison with the existing studies [5,13,25,33] on bias evaluation of gender classifier. As a fair model is supposed to have consistent accuracies across all subgroups, this implies that a fair model would have a max-min accuracy ratio closer to 1 and a DoB closer to 0.

For gender-independent facial attribute classifiers, following the studies in [3,27,36], we used classification accuracy, True Positive Rate (TPR), Difference of Equal Opportunity (DEO) and Difference in Equalized Odds (DEOdds) where Equal Opportunity (EO) requires a classifier to have equal TPRs on each subgroup and a violation of this equal opportunity is measured by the DEO. DEOdds measures the absolute difference in the probability of correctly predicting the positive class between the subgroups for each actual outcome, summed over all possible outcomes [3,27,36]. Furthermore, we also analyzed the maximum group accuracy and the minimum group accuracy associated with the best and worst performing demographic subgroups, respectively. A model that can maintain or improve accuracy and TPR while reducing DEO and DEOdds would be an ideal classifier in terms of enhancing accuracy as well as fairness.

5 Results

5.1 Face-Based Gender Classification

In this section, we will discuss the performance and fairness of the face-based gender classifier across gender-racial groups.

Intra-dataset Evaluation: Table 2 shows the performance of the baseline ResNet50 model and our proposed FineFACE model in gender classification when trained and tested on the FairFace dataset. As can be seen, our proposed FineFACE model reduced the Degree of Bias (DoB) and the Max-Min accuracy ratio by approximately 86% and 13%, respectively, over the baseline. At the same time, the overall classification accuracy improved by about 3% over the baseline ResNet50 model. Note, we also evaluated and compared performance of the baseline DenseNet architecture over FineFACE using DenseNet backbone for gender classification. The experimental results demonstrated the efficacy of FineFace in improving accuracy and fairness over the baseline DenseNet architecture. Thus, highlighting the importance of systematic construction of experts from shallow to deep layers followed by attention region prediction and multistage learning by FineFACE over feature reuse between shallow to deep layers by the baseline DenseNet. More details on the experimental results are given in Section 1.1 of the supplementary material.

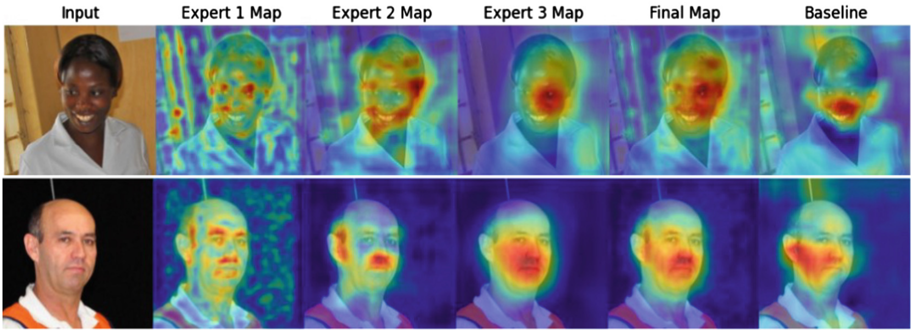


Fig. 3. Visualization Results of Gender Classifier. Left through right in each set of images are the input image from FairFace dataset, visualization results based on our FineFACE method's 3 experts ($\tilde{\Omega}_1^{norm}$, $\tilde{\Omega}_2^{norm}$, $\tilde{\Omega}_3^{norm}$), and our method's final visualization ($\tilde{\Omega}_{oval}^{norm}$), versus a basic ResNet50 architecture's final visualization ($\tilde{\Omega}_{ori}^{norm}$). Our FineFACE captures a more comprehensive feature representation of the image, thereby enhancing fairness as well as accuracy.

Cross-Dataset Evaluation: Table 3 shows the results of the baseline ResNet50 and our proposed FineFACE, based on ResNet50 backbone when trained on FairFace and evaluated on UTKFace and LFWA+ datasets. Our model significantly reduced the bias of the gender classifier by reducing DOB by approximately 55% and 77%, and Max-Min ratio by 18% and 17% over the baseline even on the cross-dataset evaluation, respectively. Overall, performance degradation of the classifiers is minimal on cross-dataset evaluation except for the UTKFace dataset due to poor quality samples majorly showing age progression. Table 4 shows the results on the CelebA test set. Our model reduced DOB by approximately 41% and Max-Min ratio by 2%. These results demonstrate the efficacy of our model in significantly reducing bias as well as improving accuracy even on the cross-dataset evaluation.

Figure 3 shows the visualization of the attention map learned by the ResNet50-based FineFACE and the baseline ResNet50-based gender classifier. For proposed FineFACE, we generate 4 heatmaps for each image, i.e., $\tilde{\Omega}_1^{norm}$ from expert 1, $\tilde{\Omega}_2^{norm}$ from expert 2, $\tilde{\Omega}_3^{norm}$ from expert 3 and $\tilde{\Omega}_{oval}^{norm}$ which is the aggregation of the three experts' heatmaps and is used for the final prediction (refer to Sect. 3.2). The maps generated by Expert 1 show a focus on low-level features such as edges, evident from the scattered attention across the face, capturing details such as the outline of the face, eyes, nose, and mouth. The maps generated by Expert 3 have more concentrated attention on key facial regions that are critical for gender classification, such as the central face area. Thus, there is a clear progression in the focus of attention from Expert 1 to Expert 3 and all the varying levels of attention are captured in the final concatenated map (Final Map). As the original ResNet50 has only a 1 classifier for prediction, we generated 1 heatmap $\tilde{\Omega}_{ori}^{norm}$ using the feature maps from the last convolutional layer. Note, $\tilde{\Omega}_3^{norm}$ and $\tilde{\Omega}_{ori}^{norm}$ are both generated based on the feature maps of the last convolutional layer of the ResNet50 backbone, but $\tilde{\Omega}_3^{norm}$ captures much more comprehensive and accurate information than $\tilde{\Omega}_{ori}^{norm}$. Further, the overall feature map from the proposed FineFACE model illustrates the efficacy of the fine-grained framework in capturing comprehensive and discriminant regions vital for gender classification over the baseline.

Table 2. Gender Classification Accuracy (%) on FairFace testset across different demographics using baseline ResNet50 and our proposed FineFACE. M stands for male and F stands for female. The top performance results are highlighted in bold.

Race	White		Black		East Asian		SE Asian		Latino		Indian		Middle E				
Gender	M	F	M	F	M	F	M	F	M	F	M	F	M	F	Max/ Min↓	Overall↑	DoB↓
Baseline	96.5	89.9	94.4	82.4	97.2	88.9	94.4	91.5	95.6	92.2	98.1	93.3	97.8	92.4	1.18	93.2	4.2
FineFACE	97.1	97	97.2	96.2	97	96.2	96.2	97	96	96.3	96.6	95.9	96.1	95.1	**1.02**	**96.4**	**0.6**

Comparison with Published Work: Table 5 shows the performance of our proposed FineFACE method over published bias mitigation techniques based on multi-tasking [5], adversarial debiasing [33], deep generative views [25], and consistency regularization [13]. All these studies are reported for the ResNet50 based gender classifier trained and tested on the FairFace dataset.

As can be seen, our proposed FineFACE obtained the lowest DoB of 0.26 and the Max-Min accuracy ratio of 1.008 over all the existing published studies. Moreover, the overall accuracy was not only maintained but also increased by 1.74% compared to the second-best model (indicated as D in the Table) based on Deep generative views [25]. Therefore, our proposed method obtains state-of-the-art performance.

Table 3. Cross-dataset evaluation - Gender Classification Accuracy (%) on UTKFace and LFWA+ test sets across different demographics for baseline ResNet50 and our proposed FineFACE.

Dataset	Race	White		Black		Asian		Indian		Others/Undefined		Max/Min↓	Overall↑	DoB↓
	Gender	M	F	M	F	M	F	M	F	M	F			
UTKFace	Baseline	90.2	72.2	94.6	67.6	88.2	65.7	95.1	74.9	89.1	78.8	1.45	81.9	11.1
	FineFACE	91	86.5	93.9	79.7	91.1	81.1	95.1	86.3	90	88.3	**1.19**	**88.5**	**5**
LFWA+	Baseline	96.9	89.1	98.7	78.8	96.5	78.3	97.9	95	96.8	91.1	1.26	95.3	7.7
	FineFACE	99.1	98	98.9	95.3	98.6	94.8	100	100	98.4	96.2	**1.05**	**98.6**	**1.9**

Table 4. Cross-dataset evaluation - Gender Classification Accuracy (%) on CelebA testset across gender for baseline Resnet50 and our proposed FineFACE.

Gender	M	F	Max/Min↓	Overall↑	DoB↓
Baseline	90.6	94.9	1.05	93.2	2.2
FineFACE	96.4	99	**1.03**	**98**	**1.3**

Further, existing bias mitigation techniques based on adversarial debiasing [33] and multitasking [5] need demographically annotated data during training. The generative techniques based on deep generative views [25] and consistency regularization [13] are computationally very expensive and obtain low generalizability. Compared to the existing methods, the proposed FineFACE offers significant advantages: it mitigates bias in the absence of protected attributes, offers high generalizability, and is application-agnostic. Importantly, our method significantly improves fairness along with overall classification accuracy, emphasizing the importance of fine-grained classification.

Table 5. Comparative Analysis with FineFACE. A: Multi-Tasking [5], B: Adversarial debiasing [33], C: Consistency Regularization [13] D: Deep Generative Views based [25]. The top performance results are highlighted in bold.

Method	Accuracy								DoB↓	Max/Min↓
	Black	East Asian	Indian	Latino Hispanic	Middle Eastern	Southeast Asian	White	Overall↑		
A	91.26	94.45	95.05	95.19	97.35	94.2	94.96	94.64	1.81	1.067
B	87.66	91.93	93.67	93.8	95.96	91.81	93.96	92.69	2.62	1.095
C	90.83	93.6	94.48	94.7	95.94	93.64	94.57	94	1.59	1.056
D	91.64	95.29	95.38	95.32	97.11	93.5	94.92	94.72	1.72	1.06
FineFACE	96.21	96.84	96.37	96.61	96.53	96.04	96.55	**96.46**	**0.26**	**1.008**

5.2 Gender-Independent Facial Attribute Classification

In this section, we will discuss the performance of our 13 gender-independent facial attribute classifiers. We report mean scores over the 13 labels [27] called gender-independent target attribute (refer to Sect. 4 for more details on the 13 target attributes) with gender as the protected attribute.

Table 6 shows the performance of the gender-independent facial attribute classifier using our proposed FineFACE over the baseline ResNet50 model. FineFACE improves overall accuracy, minimum group accuracy, and TPR, while significantly reducing bias by approximately 91% (DEO) and 92% (DEODD). There is a marginal reduction in maximum group accuracy by only 1.52%. Worth mentioning, facial attributes like "bags under eyes", "chubby", "high cheekbones", "narrow eyes", and "smiling" are more subtle in nature. Thus more detailed features from lower layers help in understanding the subtle cues differentiating a normal cheekbone from a high cheekbone, a narrow eye from a normal eye, and a natural curvature of lips versus a smile across gender. Thus, obtaining performance enhancement as well as bias reduction for these gender-independent facial attributes.

Table 6. Facial Attribute Classification Accuracy (%) on the CelebA dataset for baseline and our proposed FineFACE - mean scores over the 13 attributes [27] called gender-independent are reported. The top performance results are highlighted in bold.

Method	Accuracy↑	Max. grp. Acc.	Min. grp. Acc.	TPR↑	Max. grp. TPR	Min. grp. TPR	DEO↓	DEODD↓
Baseline	92.47	**94.46**	90.14	67.9	73.88	61.34	12.54	16.54
FineFACE	**92.85**	92.94	**92.76**	**76.48**	**76.97**	**75.79**	**1.18**	**1.38**

Comparison with Published Work: In this section, we compare the performance of our proposed FineFACE over bias mitigation techniques namely, domain-independent models [29] (Domain Indep.), regularization [24,30] (Regularized), FairMixup [3], GAN-based offline dataset debiasing [27] (GAN Debiasing), and adaptive sampling [36] (g-SMOTE + Adaptive Sampling), reported for the 13 gender-independent facial attribute classifiers.

We summarized the results in the Table 7. The proposed FineFACE has achieved improved performance in both overall accuracy and accuracy of the worst-performing group compared to the baseline classifier. Although there is a slight reduction in the accuracy of the best-performing group (by 1.52%), the performance of this group remains comparable to the overall and worst-performing groups which improve by approximately 0.4% and 2.5% respectively. Worth mentioning, among 6 existing bias mitigation techniques in Table 7, only g-SMOTE [27] and g-SMOTE adaptive sampling [36] are able to improve performance of the worst performing group, with respect to the baseline, along with our proposed FineFACE, thus obtaining **pareto-efficiency**. Furthermore, our

Table 7. Fairness methods on the CelebA dataset - mean scores over the 13 labels [27] called gender independent. The top performance results are highlighted in bold.

Method	Weighting*	Domain Indep.*	Baseline single task	GAN Debiasing	Regularized	g-SMOTE + Adap. Sampl.	g-SMOTE	Baseline FairMixup	Fair Mixup	FineFACE [ours]
Accuracy↑	91.45	91.24	92.47	92.12	91.05	92.56	92.64	92.74	88.46	**92.85**
Max. grp. Acc.	93.35	93.04	94.46	94.03	94.42	94.44	**94.59**	93.85	90.42	92.94
Min. grp. Acc	89.06	88.93	88.93	89.85	87.86	90.36	90.35	91.44	86.36	**92.76**
TPR↑	64.02	70.74	67.90	66.13	54.2	67.11	66.14	**79.13**	46.67	76.48
Max. grp. TPR	67.41	75.61	73.88	70.36	56.11	74.06	73.43	**80.89**	47.85	76.97
Min. grp. TPR	59.74	66.05	61.34	61.25	52.34	59.78	58.32	72.92	44.27	**75.79**
DEO↓	7.67	9.56	12.54	9.11	3.77	14.28	15.11	7.97	3.58	**1.18**
DEODD↓	9	13.29	16.54	12.04	5.06	19.3	19.32	10.06	4.29	**1.38**

Table 8. Minimum Group Accuracy for the 13 gender-independent individual attributes.

Attribute Name	Baseline	FineFACE
Bags Under Eyes	73.24	80.73
Bangs	94.67	95.9
Black Hair	86.4	90.51
Blond Hair	91.96	94.41
Brown Hair	81.26	89.14
Chubby	89.29	95.64
Eyeglasses	99.23	99.73
Gray Hair	95.28	98.35
High Cheekbones	85.53	87.83
Mouth Slightly Open	93.46	94.23
Narrow Eyes	91.97	87.99
Smiling	91.64	93.17
WearingHat	98.21	99.08

method obtains the highest improvement in the TPR of the worst-performing groups, along with the second-best results for overall TPR and the TPR for the best-performing groups. The **key highlight** of our method is the substantial reduction in both DEO and DEOdds, 3× lower than the next best method i.e., Fair Mixup [3]. Comparison of the various fairness methods is visually represented in Fig. 1, Sect. 1.2 of the supplementary material

We also reported the minimum group accuracy of the baseline ResNet50 model and our FineFACE for each of the 13 gender-independent attributes individually in Table 8. Our model outperformed the baseline for all except 1 attributes ("Narrow Eyes" by the baseline is more accurate by approximately 4%). For the other 12 attributes, our model outperformed at least by 0.5% and at most by 7.9%. Thus, we also demonstrate the efficacy of our FineFACE in improving the minimum group accuracy for the majority (12 out of 13 attributes) of facial attributes on an individual basis.

6 Conclusion

The task of facial attribute classification presents inherent complexities due to high inter-class similarity, significant intra-class variation, and demographic diversity, which often result in performance disparities across protected attributes.

To effectively tackle these challenges, it is essential to incorporate local and subtle cues into the classification process. In our research, we propose a novel fine-grained feature framework designed for demographically fair facial attribute classification. This framework integrates detailed low-level and semantic high-level information across shallow to deep layers of the model. Through extensive evaluation on widely used facial attribute datasets, our approach demonstrates significant effectiveness in learning fair representation, achieving up to a three-fold reduction in bias compared to state-of-the-art bias mitigation techniques. Importantly, our method achieves a Pareto-efficient balance between accuracy and fairness without requiring the presence of protected attribute labels during classifier training—a critical advantage given privacy concerns and regulatory constraints that often prohibit the collection of such sensitive data. Furthermore, our study marks the first benchmark evaluation of the fairness of facial attribute classifiers using fine-grained features compared to existing supervised bias mitigation techniques.

While the multi-step training strategy extends the training duration compared to the original backbone networks, the training is an offline process and the more significant concern in real-world applications is the inference cost which is affordable for our method. As part of future work, we will also explore other backbone architectures such as Transformer. In addition, we will further analyze biases across intersectional groups, such as gender + target attribute, following insights from recent studies [3,36].

Acknowledgements. This work is supported by National Science Foundation (NSF) award no. 2129173.

References

1. Albiero, V., et al.: Analysis of gender inequality in face recognition accuracy. In: Proceedings of IEEE WACV Workshops 2020, pp. 81–89 (2020)
2. Buolamwini, J., Gebru, T.: Gender shades: intersectional accuracy disparities in commercial gender classification. In: FAT. PMLR, vol. 81, pp. 77–91. PMLR (2018)
3. Chuang, C.Y., Mroueh, Y.: Fair mixup: fairness via interpolation. arXiv preprint arXiv:2103.06503 (2021)
4. Cui, J., Zhu, B., Wen, X., Qi, X., Yu, B., Zhang, H.: Classes are not equal: an empirical study on image recognition fairness. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 23283–23292 (2024)

5. Das, A., Dantcheva, A., Brémond, F.: Mitigating bias in gender, age and ethnicity classification: a multi-task convolution neural network approach. In: ECCV Workshops (1), vol. 11129, pp. 573–585 (2018)
6. Fu, J., Zheng, H., Mei, T.: Look closer to see better: recurrent attention convolutional neural network for fine-grained image recognition. In: Proceedings of IEEE Conference on CVPR, pp. 4438–4446 (2017)
7. Gao, Y., Han, X., Wang, X., Huang, W., Scott, M.: Channel interaction networks for fine-grained image categorization. In: AAAI Conference, pp. 10818–10825 (2020)
8. Guo, G., Zhang, N.: A survey on deep learning based face recognition. Comput. Vis. Image Underst. **189** (2019)
9. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on CVPR, pp. 770–778 (2016)
10. Jiang, P.T., Zhang, C.B., Hou, Q., Cheng, M.M., Wei, Y.: Layercam: exploring hierarchical class activation maps for localization. IEEE Trans. IP **30**, 5875–5888 (2021)
11. Karkkainen, K., Joo, J.: Fairface: face attribute dataset for balanced race, gender, and age for bias measurement and mitigation. In: Proceedings of IEEE/CVF Winter Conference on Applications of Computer Vision, pp. 1548–1558 (2021)
12. Krishnan, A., Almadan, A., Rattani, A.: Understanding fairness of gender classification algorithms across gender-race groups. In: Proceedings of 19th IEEE ICMLA, pp. 1028–1035 (2020)
13. Krishnan, A., Rattani, A.: A novel approach for bias mitigation of gender classification algorithms using consistency regularization. Image Vis. Comput. **137**, 104793 (2023)
14. Lin, T.Y., RoyChowdhury, A., Maji, S.: Bilinear convolutional neural networks for fine-grained visual recognition. IEEE Trans. PAMI **40**(6), 1309–1322 (2017)
15. Lin, X., Kim, S., Joo, J.: Fairgrape: fairness-aware gradient pruning method for face attribute classification. In: ECCV (13), vol. 13673, pp. 414–432 (2022)
16. Liu, D., Zhao, L., Wang, Y., Kato, J.: Learn from each other to classify better: cross-layer mutual attention learning for fine-grained visual classification. Pattern Recogn. **140**, 109550 (2023)
17. Liu, Z., Luo, P., Wang, X., Tang, X.: Deep learning face attributes in the wild. In: Proceedings of International Conference on Computer Vision (ICCV) (2015)
18. Liu, Z., Luo, P., Wang, X., Tang, X.: Large-scale celebfaces attributes (celeba) dataset (2018). Accessed 15 Aug 2018
19. Loshchilov, I., Hutter, F.S.: Stochastic gradient descent with warm restarts (2016)
20. Luo, W., Zhang, H., Li, J., Wei, X.S.: Learning semantically enhanced feature for fine-grained image classification. IEEE SPL **27**, 1545–1549 (2020)
21. Majumdar, P., Singh, R., Vatsa, M.: Attention aware debiasing for unbiased model prediction. In: Proceedings of IEEE/CVF ICCVW 2021, pp. 4116–4124 (2021)
22. Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., Galstyan, A.: A survey on bias and fairness in machine learning. ACM Comput. Surv. **54**(6) (2021)
23. Muthukumar, V.: Color-theoretic experiments to understand unequal gender classification accuracy from face images. In: Proceedings of IEEE CVPR Workshops 2019, pp. 2286–2295 (2019)
24. Padala, M., Gujar, S.: FNNC: achieving fairness through neural networks. In: Proceedings of IJCAI 2020, pp. 2277–2283. ijcai.org (2020)
25. Ramachandran, S., Rattani, A.: Deep generative views to mitigate gender classification bias across gender-race groups. In: Proceedings of ICPR 2022 International Workshops and Challenges, vol. 13645, pp. 551–569. Springer, Cham (2022)

26. Ramachandran, S., Rattani, A.: A self-supervised learning pipeline for demographically fair facial attribute classification. In: The IEEE International Joint Conference on Biometrics (IJCB). IEEE (2024)
27. Ramaswamy, V.V., Kim, S.S.Y., Russakovsky, O.: Fair attribute classification through latent space de-biasing. In: CVPR, pp. 9301–9310 (2021)
28. Wang, Q., Xie, J., Zuo, W., Zhang, L., Li, P.: Deep CNNs meet global covariance pooling: better representation and generalization. IEEE Trans. PAMI **43**(8), 2582–2597 (2020)
29. Wang, Z., et al.: Towards fairness in visual recognition: effective strategies for bias mitigation. In: Proceedings of IEEE/CVF CVPR 2020, pp. 8916–8925 (2020)
30. Wick, M.L., Panda, S., Tristan, J.: Unlocking fairness: a trade-off revisited. In: Advances in Neural Information Processing Systems (NeurIPS 2019), vol. 32, pp. 8780–8789 (2019)
31. Zafeiriou, S., Zhang, C., Zhang, Z.: A survey on face detection in the wild: past, present and future. Comput. Vis. Image Underst. **138**, 1–24 (2015)
32. Zeiler, M.D., Fergus, R.: Visualizing and understanding convolutional networks. In: ECCV, pp. 818–833 (2014)
33. Zhang, B.H., Lemoine, B., Mitchell, M.: Mitigating unwanted biases with adversarial learning. In: Proceedings of AAAI/ACM AIES 2018, pp. 335–340. ACM (2018)
34. Zhang, Z., Song, Y., Qi, H.: Age progression/regression by conditional adversarial autoencoder. In: Proceedings of IEEE Conference on CVPR (2017)
35. Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., Torralba, A.: Learning deep features for discriminative localization. In: Proceedings of IEEE Conference on CVPR, pp. 2921–2929 (2016)
36. Zietlow, D., et al.: Leveling down in computer vision: pareto inefficiencies in fair deep classifiers. In: Proceedings of IEEE/CVF Conference on CVPR, pp. 10410–10421 (2022)

One-Factor Cancelable Biometric Template Protection Scheme for Real-Valued Features

Ruoqi Zhang[1,2,3], Peisong Shen[1,3](✉), Kewei Lv[1,2,3], and Chi Chen[1,2,3]

[1] Institute of Information Engineering, Chinese Academy of Sciences, Beijing, China
{zhangruoqi,lvkewei,chenchi}@iie.ac.cn
[2] School of Cyber Security, University of Chinese Academy of Sciences, Beijing, China
[3] Key Laboratory of Cyberspace Security Defense, Beijing, China
shenpeisong@iie.ac.cn

Abstract. Biometric template protection is crucial in biometric recognition. Cancelable biometrics, a method for template protection, typically requires both human biometric data and a token as inputs for template generation. However, when the token is compromised, the template is vulnerable to various security attacks, leading to privacy breaches. This paper focuses on one-factor cancelable biometric scheme that relieves users from the burden of managing tokens. Specifically, we propose a new one-factor cancelable biometric scheme for protecting real-valued biometric features. It first obtains a hashed code from the biometirc feature using prime Cosine LSH functions, then generates an index sequence from the hashed code, finally, reorders a random string using the index sequence to create a cancelable template. The experiment results demonstrate the efficiency of the proposal scheme on real-valued facial, ear and fingerprint biometric features. We also verify the unlinkability, revocability and irreversibility of our scheme against various attacks.

Keywords: Cancelable biometrics · Template Protection · Locality Sensitive Hashing · Generated-Index-based Reordering

1 Introduction

In recent years, biometric recognition has been widely used for identity verification. However, once a biometric feature is leaked, the user's identity privacy is permanently compromised, leading to security and privacy issues. Therefore, biometric template protection (BTP) is crucial for privacy.

BTP schemes are categorized into two types: cancelable biometrics (CB) and biometric cryptosystems (BC), with the former being the focus of this paper. Cancelable biometrics generate protected templates from the original biometric features through an irreversible transformation with user-specific parameters. These templates are then matched within the transformation domain, ensuring

A. Antonacopoulos et al. (Eds.): ICPR 2024, LNCS 15329, pp. 98–115, 2025.
https://doi.org/10.1007/978-3-031-78110-0_7

the protection of the original data. It is imperative that a CB scheme fulfills the following four criteria [26]:

- Irreversibility: It is computationally infeasible to reconstruct the original feature from a protected template.
- Revocability: The capacity to revoke an old protected template and generate a new one as required, with no correlation between two templates.
- Unlinkability: The inability to determine whether two or more protected templates belong to the same user, preventing user identity leakage caused by cross-matching.
- Performance preservation: The recognition accuracy of the protected template should be comparable to that of the original plain template.

1.1 Two-Factor Cancelable Biometrics

In 2001, Ratha and Connell [29] proposed the core idea of cancelable biometric template protection. Generally, cancelable biometrics are designed as parameterized schemes, requiring the user to provide both the biometric feature and a token (as a key), are frequently referred to as "two-factor" or "tokenized".

A well-known instance is Biohashing [13]. Through the inner product operation between the feature and user-specific vectors, followed by binarization, a protected template is generated. Nevertheless, if the tokens (i.e., the user-specific vectors) are stolen, the pre-image can be easily obtained from the template, leading to illegal access [24], as well as linkage and intrusion attacks [3].

The implementation of cancelable biometrics based on Random Projection (RP) [28] reduces the dimensionality of the feature through random Gaussian matrices to achieve irreversibility. However, RP schemes can recover the biometric features when multiple tokens and templates are leaked, leading to privacy breaches [5]. To solve this problem and improve cancelable biometrics security, a secure method called Absolute Value Equations Transform (AVET) has been proposed [5].

In recent years, Locality Sensitive Hashing (LSH) has emerged as a significant method for cancelable biometrics. LSH has been used for generating cancelable biometrics for iris [31] and fingerprint [30], based on the Hamming metric space. A ranking-based LSH method, termed "Index-of-Max" (IoM) hashing, has been proposed to inspired cancelable fingerprint templates, with two implementations derived from the IoM hashing concept: GRP-IoM and URP-IoM [14]. Lai et al. [19] introduced cancelable iris templates based on Indexing-First-Order (IFO) hashing. Jiang et al. [12] have implemented cancelable face templates based on LSH under Euclidean and Cosine metrics. However, utilizing Integer Linear Programming and Quadratic Constrained Quadratic Programming, attackers can impersonate legitimate uses and conduct reversible attacks on the URP-IoM [7]. If tokens are stolen, the unlinkability and irreversibility of the IoM scheme will be compromised, making it vulnerable to illegal access [9].

1.2 One-Factor Cancelable Biometrics

One-factor cancelable biometrics, also known as tokenless cancelable biometrics, requires only the user's biometrics, without tokens.

It was first introduced in [27] for protecting iris codes. The scheme extracts consistent bits from multiple iris codes of a user, groups them into fixed-length address words $\boldsymbol{S}$. Using $\boldsymbol{S}$ to address a random bitstring $\boldsymbol{r}$ generates a template $\boldsymbol{t}$, $\boldsymbol{r}$ is also stored. However, this scheme is invertible when $\boldsymbol{r}$ is known to an adversary, revealing that $\boldsymbol{r}$ must be kept secret [18].

Inspired by the IFO scheme [19], a one-factor cancelable biometrics for binary fingerprint authentication has been developed [16]. During enrollment, the scheme transforms the random string $\boldsymbol{r}$ using IFO to obtain a pseudo-identifier $\boldsymbol{PI}$, which is stored as a template. The biometric feature is transformed through IFO to obtain a hashed code $\boldsymbol{h}$, which is then XORed with $\boldsymbol{r}$ to obtain ciphertext $\boldsymbol{c}$. The stored $\boldsymbol{c}$ is used to protect $\boldsymbol{r}$. During authentication, the biometric feature is used to recover $\boldsymbol{r'}$ from $\boldsymbol{c}$. The pseudo-identifier $\boldsymbol{PI'}$ is generated from $\boldsymbol{r'}$ using the same transformation as in the enrollment stage, and the similarity between $\boldsymbol{PI}$ and $\boldsymbol{PI}'$ is compared. Wang and Li [32] enhanced the GRP-IoM scheme [14] to achieve one-factor cancelable palmprint templates. The enrollment and authentication of $\boldsymbol{PI}$ and $\boldsymbol{c}$ are similar to those in [16]. In [32], the $\boldsymbol{PI}$ is generated by applying minimum signature hashing to $\boldsymbol{r}$. And $\boldsymbol{h}$ is generated using GRP-IoM with orthogonal Gaussian projection matrices, referred to as OIOM. However, in these two one-factor cancelable biometirc schemes, during template enrollment, the generated pseudo-identifiers $\boldsymbol{PI}$ are essentially unrelated to the biometric features. The biometric features in these schemes are primarily used for the secure storage and to recover the random string $\boldsymbol{r}$.

Another one-factor cancelable scheme, extended feature vector hashing (EFV), has been introduced to protect fingerprint features [22]. It takes a binary fingerprint vector as input, generates a permutation seed sequence from the extended feature vector, and permutes a random binary string $\boldsymbol{r}$ with it to produce a cancelable template. The extended feature vector is stored after being XORed with $\boldsymbol{r}$ as auxiliary data, ensuring that neither $\boldsymbol{r}$ nor the biometric data are stored. This method is also utilized in multimodal cancelable biometric authentication schemes [21,23]. Some scholars have proposed improvements to the extended feature vector. A one-factor cancelable fingerprint template protection was proposed through minimum hash signature and a secure extended feature vector [25]. By extending the generation of permutation sequences for EFV, a one-factor cancelable template, termed "indexing self-coding", was developed to protect binary fingerprint features [8]. Nevertheless, the original fingerprint vectors can be revealed through correlation analysis between multiple auxiliary data and protected templates of the EFV scheme [22].

1.3 Motivations and Contributions

Currently, token-based two-factor CB schemes require users to retain or carry a user-special token and input it alongside their biometric data during authentication. This method poses security risks if the token is stolen or leaked [3,7,9,24].

For one-factor, i.e., tokenless cancelable biometrics authentication, the most existing research focuses on binary features [8,16,22,25,27], leaving a significant gap in the study of one-factor real-valued template protection. Additionally, in some one-factor schemes, the enrollment template is unrelated to the biometric features [16,32]. Furthermore, one-factor schemes face several security issues, for instance, EFV-based one-factor schemes [8,21,22] are vulnerable to multiple-records attacks, and none of the known one-factor schemes address resistance to similarity-based attacks.

This paper presents a secure one-factor cancelable biometrics authentication method tailored for real-valued features, with the following key contributions:

- This paper presents a new one-factor cancelable template protection scheme for real-valued biometric features. It generates a hashed code from the biometric feature using LSH under Cosine metric. Then an index sequence is generated from the hashed code, and used to reorder a random string to produce protected template. For the generated template, we prove that each bit of the random string has an equal probability of selection, enhancing privacy protection and reducing the impact of noise on recognition accuracy.
- This scheme focuses on real-valued biometric features. Experiments are conducted on facial, ear, and fingerprint datasets, i.e. CASIA-FaceV5, IITD-E and FVC2002 DB3. The experimental results demonstrate that the proposed scheme is applicable to various real-valued biometric features.
- This paper comprehensively evaluates the security and privacy of the proposed scheme from both theoretical and experimental perspectives. It analyzes the irreversibility, unlinkability, and revocability of the proposed scheme, and its resistance to multiple privacy and security attacks. In addition to common attacks, the resistance to similarity-based attacks is also evaluated. The results show that the proposed one-factor cancelable biometric scheme satisfies the criteria for template protection schemes and has significant application prospect.

The structure of the article is organized as follows: Sect. 2 describes the preliminary aspects of this paper. In Sect. 3, we introduce the proposed one-factor cancelable biometric scheme. Section 4 presents the experimental results from facial, ear and fingerprint databases. Section 5 presents a detailed security and privacy analysis of the proposed scheme. Finally, the conclusions are presented in Sect. 6.

Table 1. Notations and their descriptions

Notations	Descriptions
$\boldsymbol{f}$	Feature vector
$\mathcal{H}$	Locality Sensitive Hashing (LSH) family
$h(\cdot)$	Locality Sensitive Hashing function $h(\cdot) \in \mathcal{H}$
m	Number of LSH functions, a prime number
$\boldsymbol{u}$	Cosine LSH-based hashed code
k	Length of subcode, $2^k < m$
$\dot{\boldsymbol{u}}$	The concatenation of k repetitions of $\boldsymbol{u}$
$\hat{\boldsymbol{u}}_i$	The ith subcode of $\dot{\boldsymbol{u}}$
$\bar{u}_i$	The decimal representation of $\hat{\boldsymbol{u}}_i$
$\boldsymbol{q}$	Generated-Index sequence
$\boldsymbol{r}$	Random bitstring
$\boldsymbol{t}$	Protected template
$\boldsymbol{b}$	Encrypted bitstring
τ	Threshold of matching
S	Similarity score between protected templates

2 Preliminaries

This section introduces locality sensitive hashing (LSH) method for the cosine metric. The notations used in this paper are summarized in Table 1.

Locality Sensitive Hashing (LSH) was initially designed to solve the approximate nearest neighbor search problem [11]. Points that are closer in the original metric space have higher collision probability in the hash domain, and vice versa. The LSH function for the cosine metric [4] is defined as in Eq.(1).

$$h_{\boldsymbol{a}}(\boldsymbol{u}) = \begin{cases} 1, & if\ \boldsymbol{a} \cdot \boldsymbol{u} \geq 0 \\ 0, & if\ \boldsymbol{a} \cdot \boldsymbol{u} < 0 \end{cases} \tag{1}$$

where a random vector $\boldsymbol{a}$ is selected from the n-dimensional Gaussian distribution (i.e., each coordinate is drawn from the 1-dimensional Gaussian distribution). Using m Cosine LSH functions, it is possible to transform an n-dimensional real-valued vector into a m-length binary string.

3 Methodology

This section begins with an overview of our scheme in this paper in Fig. 1, followed by a detailed description. The feature extraction algorithm used in this paper, which extracts a fixed-length real-valued feature $\boldsymbol{f}$ from biometric data, is introduced in Sect. 4.

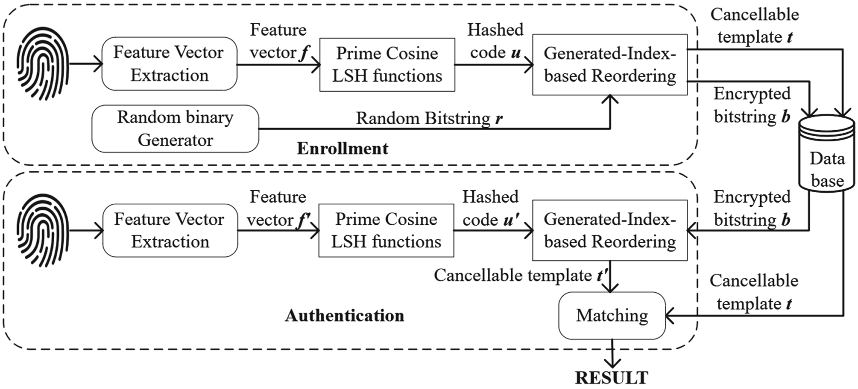

Feature Vector Extraction
Feature vector $\boldsymbol{f}$
Prime Cosine LSH functions
Hashed code $\boldsymbol{u}$
Generated-Index-based Reordering
Cancellable template $\boldsymbol{t}$
Encrypted bitstring $\boldsymbol{b}$
Random binary Generator
Random Bitstring $\boldsymbol{r}$
Enrollment
Data base
Feature Vector Extraction
Feature vector $\boldsymbol{f'}$
Prime Cosine LSH functions
Hashed code $\boldsymbol{u'}$
Generated-Index-based Reordering
Encrypted bitstring $\boldsymbol{b}$
Cancellable template $\boldsymbol{t'}$
Cancellable template $\boldsymbol{t}$
Authentication
Matching
RESULT

Fig. 1. An overview of one-factor cancelable biometric scheme

3.1 Hashed Code Generation Based on Cosine LSH

This stage embeds the n-dimensional real-valued feature $\boldsymbol{f}$ into a m-dimensional binary hashded code $\boldsymbol{u}$ using m Cosine LSH functions, as shown in Algorithm 1.

Each hash function $h_{i=1,\cdots,m} \in \mathcal{H}$ maps the n-dimensional feature $\boldsymbol{f}$ into a binary number $\boldsymbol{u}_i$ [4]. By combining m independent hash functions, the feature is converted into a uniformly distributed binary hashed code $\boldsymbol{u} \in \mathbb{Z}_2^m$. Each LSH function is a dimensionality reduction process aimed at enhancing irreversibility. Before and after the transformation, the pairwise distance relationship between vectors is maintained.

Algorithm 1. Hashed Code Generation: Cosine LSH-based

```
Input: Feature vector $\boldsymbol{f} \in \mathbb{R}^n$, $m$ Cosine LSH functions $h_{i=1,\cdots,m} \in \mathcal{H}$, $m$ is a prime
Output: binary hashed code $\boldsymbol{u} \in \mathbb{Z}_2^m$
for $i = 1$ to $m$ do
    $\boldsymbol{u}_i = h_i(\boldsymbol{f})$
end for
return $\boldsymbol{u}$
```

3.2 Cancellble Template Generation Based on Generated-Index-Based Reordering (GI-R)

This stage employs our proposed Generated-Index-based Reordering (GI-R) method to generate the protected template $\boldsymbol{t}$ from $\boldsymbol{u}$, as shown in Algorithm 2.

The key of Generated-Index-based Reordering (GI-R) is to generate an index sequence from $\boldsymbol{u}$, such that each bit of $\boldsymbol{r}$ is selected with equal probability and reordered. The implementation of GI-R is described as follows:

Algorithm 2. Protected Template Generation

```
**Input:** : Hashed code $\boldsymbol{u} \in \mathbb{Z}_2^m$, system parameter $k$, $m$, $m$ is a prime, $2^k < m$
**Output:** Protected template $\boldsymbol{t} \in \mathbb{Z}_2^m$, bitstring $\boldsymbol{b} \in \mathbb{Z}_2^m$
Generate a $m$-bitstring $\boldsymbol{r} \in \mathbb{Z}_2^m$, Initialize $\boldsymbol{b} = [0]^m$
$\dot{\boldsymbol{u}} = \underbrace{\boldsymbol{u}|\boldsymbol{u}|\cdots|\boldsymbol{u}}_{k}$, $\dot{\boldsymbol{u}} \in \mathbb{Z}_2^{km}$
for $i = 1$ to $m$ do
    $\hat{\boldsymbol{u}}_{\boldsymbol{i}} = [\dot{\boldsymbol{u}}_{(i-1)k+1}|\dot{\boldsymbol{u}}_{(i-1)k+2}|\cdots|\dot{\boldsymbol{u}}_{ik}]$
    Convert binary $\hat{\boldsymbol{u}}_{\boldsymbol{i}} \in \mathbb{Z}_2^k$ to decimal $\bar{u}_i \in \mathbb{Z}_{2^k}$
    $\boldsymbol{q}_i = ((i \times (\bar{u}_i + 1)) \bmod m) + 1$
    $\boldsymbol{b}_i = \boldsymbol{r}_i \oplus \boldsymbol{u}_i$
    $\boldsymbol{t}_i = \boldsymbol{r}_{\boldsymbol{q}_i}$
end for
return $\boldsymbol{t}$, $\boldsymbol{b}$
```

(1) Let $\dot{\boldsymbol{u}}$ denote the concatenation of k repetitions of $\boldsymbol{u}$, $\hat{\boldsymbol{u}}_{\boldsymbol{i}}$ be the ith k-subcode of $\dot{\boldsymbol{u}}$, that is, $\hat{\boldsymbol{u}}_{\boldsymbol{i}} = [\dot{\boldsymbol{u}}_{(i-1)k+1}|\dot{\boldsymbol{u}}_{(i-1)k+2}|\cdots|\dot{\boldsymbol{u}}_{ik}]$, where $|$ denote the concatenation operation.

(2) Converting subcode $\hat{\boldsymbol{u}}_{\boldsymbol{i}}$ from a binary code to a decimal number $\bar{u}_i \in \mathbb{Z}_{2^k}$, since the subcode is k bits, then the decimal $\bar{u}_i \in \{0, \cdots, 2^k - 1\}$.

(3) Get $\boldsymbol{q}_i = ((i \times (\bar{u}_i + 1)) \bmod m) + 1$, where $\boldsymbol{q}_i \in \{1, \cdots, m\}$, and can be used as an index.

(4) The protected template $\boldsymbol{t}$ is generated by reordering the random binary string $\boldsymbol{r}$ through $\boldsymbol{q}$. Then the hash code $\boldsymbol{u}$ is XORed with $\boldsymbol{r}$ to generate an encryption bitstring $\boldsymbol{b}$, which securely stores $\boldsymbol{r}$. Finally, the random binary string $\boldsymbol{r}$ is discarded, leaving only $\boldsymbol{t}$ and $\boldsymbol{b}$ stored.

In our scheme, during the template enrollment stage, each bit in the random bitstring $\boldsymbol{r}$ is selected with equal probability, as proved in Theorem 1.

Theorem 1. *In the enrollment stage, each bit in the random bitstring* $\boldsymbol{r}$ *is selected with equal probability, i.e., the generated index sequence* $\boldsymbol{q} = [\boldsymbol{q}_i | i = 1, \cdots, m]$ *ensures that for any index* $a \in \{1, \cdots, m\}$*, the probability of finding an* $i \in \{1, \cdots, m\}$ *such that* $\boldsymbol{q}_i = a$ *is uniform, where* m *is a prime number.*

Proof. In enrollment, for any feature $\boldsymbol{f} \in \mathbb{R}^n$, since the transformation parameter of each Cosine LSH function is sampled from a n-dimensional standard Gaussian distribution, then $\boldsymbol{u}_i$ is binarized using a threshold of 0 and has an equal probability of being 0 or 1. This means each bit of the generated hashed code $\boldsymbol{u}$ is independently and uniformly distributed. Therefore, in Algorithm 2, $\bar{u}_i \in \mathbb{Z}_{2^k}$ is uniformly distributed over $\{0, \cdots, 2^k - 1\}$, where $2^k < m$.

For any $a \in \{1, \cdots, m\}$, the probability that there exists an $i \in \{1, \cdots, m\}$ such that $\boldsymbol{q}_i = a$ can be expressed as:

$$\begin{aligned} &\Pr[\exists i \in \{1, \cdots, m\}, \text{ s.t. } \boldsymbol{q}_i = a] \\ =&\Pr[\exists i \in \{1, \cdots, m\}, \text{ s.t. } a = (((\bar{u}_i + 1) \times i \pmod m) + 1)] \end{aligned} \tag{2}$$

(1) When $a \neq 1$,

$$\begin{aligned}&\Pr[\exists i \in \{1, \cdots, m\},\ \text{s.t. } a = (((\bar{u}_i + 1) \times i \pmod{m}) + 1) \cap a \neq 1]\\ =&\Pr[\exists i \in \{1, \cdots, m-1\},\ \text{s.t. } (\bar{u}_i + 1) \equiv (a-1) \times i^{-1} \pmod{m}] \times \Pr[a \neq 1]\\ =&\frac{2^k}{m-1} \times \frac{1}{2^k} \times \frac{m-1}{m} = \frac{1}{m}\end{aligned} \tag{3}$$

where i^{-1} is the modular inverse of i modulo m. Since m is a prime, every $i \in \{1, \cdots, m-1\}$ has a unique inverse modulo m.

For any $\bar{u}_i \in \{0, \cdots, 2^k - 1\}$, there are 2^k possible values for $\bar{u}_i + 1$. Therefore, the size of the value domain for $(\bar{u}_i + 1) \pmod{m}$ is 2^k, then, for any $a \in \{2, \cdots, m\}$ and $i \in \{1, \cdots, m-1\}$, the probability that $(a-1) \times i^{-1} \equiv (\bar{u}_i + 1) \pmod{m}$ is $\frac{2^k}{m-1}$.

Since each $\bar{u}_i$ is uniformly distributed in $\{0, \cdots, 2^k - 1\}$ with a probability of $\frac{1}{2^k}$, therefore, for any index $a \in \{2, \cdots, m\}$, there exists an $i \in \{1, \cdots, m-1\}$ such that the probability of a being selected is $\frac{2^k}{m-1} \times \frac{1}{2^k} \times \frac{m-1}{m} = \frac{1}{m}$.

(2) When $a = 1$,

$$\begin{aligned}&\Pr[\exists i \in \{1, \cdots, m\},\ \text{s.t. } a = (((\bar{u}_i + 1) \times i \pmod{m}) + 1) \cap a = 1]\\ =&\Pr[\exists i \in \{1, \cdots, m\},\ \text{s.t. } (\bar{u}_i + 1) \times i \equiv 0 \pmod{m}] \times \Pr[a = 1]\\ =&1 \times \frac{1}{m} = \frac{1}{m}\end{aligned} \tag{4}$$

let $i = m$, then $(\bar{u}_i + 1) \times m \equiv 0 \pmod{m}$ always holds true. Therefore, when a equals 1, the probability of $a = ((\bar{u}_i + 1) \times i \pmod{m} + 1)$ is always 1, in this case, the probability of index number a being selected is $\frac{1}{m}$.

This implies that for any index number $a \in \{1, \cdots, m\}$, the probability of a being selected is $\frac{1}{m}$, leading to the conclusion that each bit in the random bitstring $\boldsymbol{r}$ is selected with equal probability.

3.3 Similarity Score Calculation and Template Matching

During the authentication, the user is required to provide his biometric features only, implementing a one-factor authentication.

Firstly, executing Algorithm 1 to obtain the hashed code $\boldsymbol{u'} \in \mathbb{Z}_2^m$ from the extracted feature $\boldsymbol{f'}$, executing steps 2–8 of Algorithm 2 to obtain the index sequence $\boldsymbol{q'} = [\boldsymbol{q'}_i | i = 1, \cdots, m]$ from $\boldsymbol{u'}$. Then, recover $\boldsymbol{r'}$ by $\boldsymbol{r'} = \boldsymbol{b} \oplus \boldsymbol{u'}$, and generate the template $\boldsymbol{t'}$ by reordering $\boldsymbol{r'}$ according to $\boldsymbol{q'}$.

The similarity score is quantified by Hamming similarity [4]:

$$S = 1 - D(\boldsymbol{t}, \boldsymbol{t'}) = 1 - \frac{dist(\boldsymbol{t}, \boldsymbol{t'})}{m}, S \in [0, 1], \boldsymbol{t} \in \mathbb{Z}_2^m, \boldsymbol{t'} \in \mathbb{Z}_2^m \tag{5}$$

where $dist(\boldsymbol{t}, \boldsymbol{t'})$ represents the Hamming distance between templates, which is the number of differing bits between the two bitstrings, m is the length of the bitstring, and $D(\cdot)$ is normalized Hamming distance.

Specifically, a larger distance indicates lower similarity score, and vice versa, when $D(\cdot) = 0$, then $S = 1$. If the similarity score S is greater than the matching threshold τ, authentication is successful; if $S < \tau$, it failed.

In this way, a one-factor cancelable template for the real-valued feature can be achieved. This approach has two advantages: (1) Since neither $\boldsymbol{r}$ nor $\boldsymbol{u}$ is stored in the database, they cannot be leaked. An attacker must guess both $\boldsymbol{r}$ and $\boldsymbol{u}$ simultaneously, which makes the scheme highly irreversible. (2) Since $\boldsymbol{u}$ is generated by a genuine user, ensuring only the genuine user can decrypt $\boldsymbol{r}$ from $\boldsymbol{b}$ to generate a correct $\boldsymbol{t}$.

4 Experimental Results

4.1 Databases

The databases and feature extraction algorithms used in this study are detailed in Table 2.

Table 2. Information of the databases

	Facial	Ear	Fingerprint
Feature extraction Algorithm	ArcFace [6]	CNN [10]	KPCA [15]
Feature Dimension	$\mathbb{R}^{512}$	$\mathbb{R}^{512}$	$\mathbb{R}^{299}$
Database name	CASIAFace-V5 [1]	IITD-E [17][a]	FVC2002 DB3 [2][b]
Number of users	500	124	100
Number of images per user	5	at least 3	5
Number of total images	2500	493	500
Number of homologous matches	5000	790	1000
Number of heterogenous matches	3118750	120488	123750

[a] The IITD-E dataset includes both raw and automatically cropped and normalized ear images. We utilized all the raw images for our study.

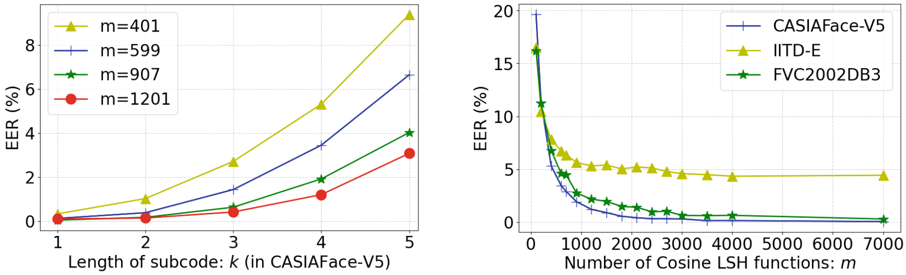
[b] The FVC2002 DB3 database comprises fingerprint images from 100 users, with each user having 8 images. For experiment, the first three images from each user are used to train the feature extraction model and the remaining five images for template transformation experiments.

4.2 Parameters of the Proposed Scheme

The cancelable biometrics require that the recognition accuracy should not significantly decline after the template transformation. This section investigates the influence of parameter selection on recognition accuracy in our scheme. Recognition accuracy is quantified using Equal Error Rate (EER), which refers to the error rate when the False Accept Rate (FAR) and False Refuse Rate (FRR) are equal. The threshold τ is set as the similarity score when FAR = FRR = EER, the lower the EER, the better the recognition performance of the system.

The two parameters used in this scheme are as follows, and shown in Fig. 2:

- Length of the subcode k when generating the index sequence.
- Number of Cosine LSH functions m, which is also the length of the template.



(a) EER with k

(b) EER with m

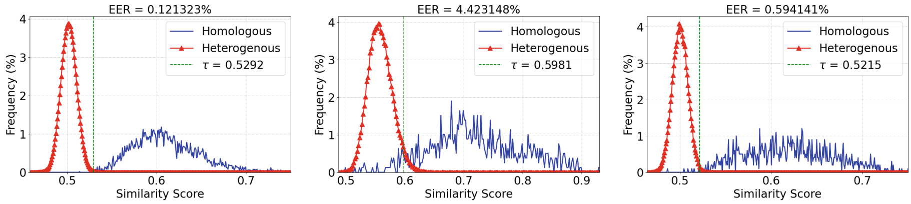
Fig. 2. EER with parameter

Effect of Parameter *k*. Figure 2a presents the curves of EER vs k in the CASIA-FaceV5 database. In the experiments, the length of the hashed code is successively fixed at 401, 599, 907, 1201. By varying the subcode length $k \in \{1, 2, 3, 4, 5\}$, experiments are conducted to determine the impact of k. It illustrates that, for a fixed value of m, the EER increases exponentially with increasing k, which is consistent with expectations. The increase k results in a higher number of consecutive bits within the hashed code being associated with each index value, making it more susceptible to noise from the hashed code. Additionally, when k remains constant, the larger m is, the lower the EER.

Effect of Parameter *m*. It aims to further verify the relationship between recognition accuracy and m. Figure 2b illustrates the curve of EER with m. Given that k is consistently set at 4, the values of m (where m is a prime number) are sequentially increased from 101, 199, 401, 599, 701, 907, 1201, 1499, 1801, 2099, 2399, 2699, 2999, 3499, 4001 and ultimately reaching 7001. The figure reveals that as the template length m increases from 101 to 1201, there is a significant reduction in EER. Subsequently, the matching accuracy improves at a slower rate and tends to stabilize when m is larger than 2999. The conclusion drawn is that an increased template length m is beneficial to improve the matching accuracy in template authentication.

4.3 Accuracy Performance Evaluation and Computation Efficiency

In this section, to validate the accuracy and efficiency of the proposed one-factor scheme, we compared it with advanced two-factor real-valued biometirc templates [5,14]. The parameters in this paper are $k = 4$, $m = 3499$. Figure 3 shows the distribution of homologous and heterogenous match similarity scores, as well as EER and threshold selection for our scheme in different databases.



(a) CASIAFace-V5 (b) IITD-E (c) FVC2002 DB3

Fig. 3. The distribution of homologous match and heterogenous match similarity scores, and threshold selection in different databases.

Table 3 presents the accuracy comparison, it can be seen that our scheme achieves higher experimental accuracy compared to two-factor schemes, except for slightly lower accuracy on fingerprint features compared to GRP-IoM. The proposed scheme performs well across three real-valued biometrics, and the EER of our scheme remains almost unchanged even after template protection.

Table 3. EER% for original features and transformed templates (lower is better)

Method	CASIAFace-V5	IITD-E	FVC2002 DB3	vector size
original	0.106253	4.431607	2.156970	$\mathbb{R}^n$
GRP-IoM [14]	0.131210	4.672924	**0.39960**	$\mathbb{Z}_{16}^{900}$
URP-IoM [14]	7.932080	7.973271	1.363030	$\mathbb{Z}_{250}^{500}$
AVET [5]	1.428473	5.206829	2.127070	$\mathbb{R}^{n/2}$
our proposal	**0.121323**	**4.423148**	0.594141	$\mathbb{Z}_2^{3499}$

The computational complexity of the scheme is evaluated by average time for template enrollment and authentication in IITD-E. The primary computational overhead in proposed scheme is from Cosine LSH and index sequence generation. Table 4 shows the average time cost on a desktop with Intel(R) Core(TM) i5-9500 CPU @ 3.00 GHz, 64-bit Windows.

The template enrollment time in the proposed scheme is superior to that in GRP-IoM. Since Cosine LSH and index sequence generation are also required for one-factor template authentication, the authentication time is approximately 8.5 milliseconds, slightly higher than in other two-factor schemes. Although the

Table 4. Average time for template enrollment and authentication in IITD-E(s)

	GRP-IoM [14]	URP-IoM [14]	AVET [5]	our proposal
Enrollment time	0.531221	0.019305	0.014317	0.049969
Authentication time	0.005943	0.007739	0.000525	0.008532

URP-IoM and AVET schemes take less time than ours, they have much lower accuracy on all three biometric databases. Our scheme emphasizes achieving high-accuracy authentication, and secure one-factor cancelable templates. Considering these factors, our one-factor cancelable biometric scheme, for real-valued biometric features, demonstrates practical value.

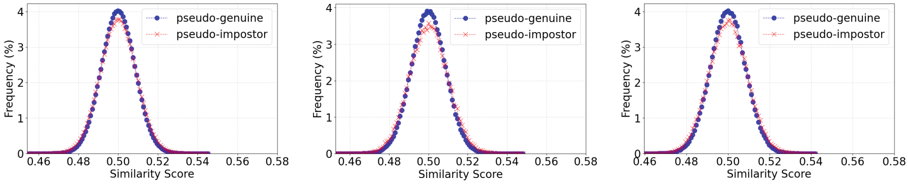
5 Security and Privacy Analysis

5.1 Unlinkability Analysis

The unlinkability of our scheme is verified by comparing the scores of pseudo-genuine and pseudo-impostor as described in [14].

The pseudo-impostor score refers to the similarity score from each template of the same feature by using different random bitstring $\boldsymbol{r}$. In the experiment, 51 random bitstrings $\boldsymbol{r}$ were used to generate 51 templates $\boldsymbol{t}$. Each user's first template is compared with the remaining 50 templates to calculate the pseudo-impostor scores for each feature. And the pseudo-genuine scores are calculated based on the templates generated from different features using different $\boldsymbol{r}$.

The experimental results of the scheme are shown in the Fig. 4, which shows that the pseudo-genuine and pseudo-impostor scores distributions are largely overlapped. Hence, the proposed scheme satisfies the unlinkability property.



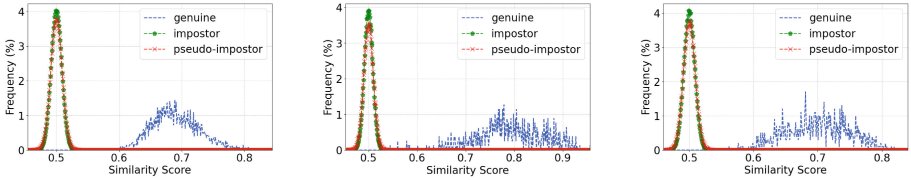
(a) CASIAFace-V5 (b) IITD-E (c) FVC2002 DB3

Fig. 4. The distributions of pseudo-genuine and pseudo-impostor scores for unlinkability analysis. Overlapped distributions indicate indistinction between templates generated from the same user or from others.

5.2 Revocability Analysis

Revocability is typically validated experimentally using three score distributions: pseudo-impostor scores, impostor scores (scores of different user's templates), and genuine scores (scores from the same user).

As shown in Fig. 5, there is a significant overlap between the pseudo-impostor and impostor score distributions. This indicates that even though the templates are generated from the same vectors, the newly generated 50 templates with the different random bitstrings $\boldsymbol{r}$ are distinctive. And the distribution of the genuine scores is distinctly separated from the impostor and pseudo-impostor scores, confirming the revocability of the proposed scheme.



(a) CASIAFace-V5 (b) IITD-E (c) FVC2002 DB3

Fig. 5. The distributions of genuine, impostor and pseudo-impostor scores for revocability analysis.

5.3 Irreversibility Analysis

Irreversibility ensures that reconstructing biometric feature $\boldsymbol{f}$ from the protected template $\boldsymbol{t}$ is computationally infeasible, thus preserving privacy even if the template is leaked. Assuming that the adversary knows the authentication algorithm and transformation parameters (e.g., $h_{i\in\{1,\cdots,m\}} \in \mathcal{H}$, m, k), for template $\boldsymbol{t} \in \mathbb{Z}_2^m$, the attacker cannot directly infer the real-valued feature $\boldsymbol{f} \in \mathbb{R}^n$ because the template $\boldsymbol{t}$ is a reordering of a random string $\boldsymbol{r}$, which is not stored.

Brute Force Attack (BFA). A brute force attack involves the adversary guessing all possible feature values directly. Assuming the worst-case scenario where the adversary knows the maximum and minimum values of the feature components and the precision, they would guess every possible value for each dimension sequentially. As shown in the Table 5, such an attack is infeasible.

Attack via Single Record. A single record attack occurs when an adversary knows a pair of the encrypted binary string $\boldsymbol{b}$ and the template $\boldsymbol{t}$ and attempts to recover $\boldsymbol{f}$. In the proposed scheme, $\boldsymbol{f}$ is first converted into a hashed code $\boldsymbol{u}$. $\boldsymbol{u}$ is then used to generate an index vector $\boldsymbol{p}$, which is used to reorder the random string $\boldsymbol{r}$ and generate a protected template $\boldsymbol{t}$. The central idea of the

Table 5. Complexity to invert single and entire feature component

	Min value with six decimal precision	Max value with six decimal precision	Possibilities for single feature component	Total possibilities for entire feature
CASIAFace-V5	−0.210062	0.214465	$424527 \approx 2^{19}$	$2^{19\times512} = 2^{9728}$
IITD-E	−0.720602	0.850098	$1570700 \approx 2^{21}$	$2^{21\times512} = 2^{10752}$
FVC2002 DB3	−0.250376	0.213155	$463531 \approx 2^{19}$	$2^{19\times299} = 2^{5681}$

single record attack is to recover $\boldsymbol{r}$ first, estimate $\boldsymbol{u}$ through $\boldsymbol{b} \oplus \boldsymbol{r}$, and finally use $\boldsymbol{u}$ along with $h_{i\in\{1,\cdots,m\}} \in \mathcal{H}$ to recover $\boldsymbol{f}$.

Since $\boldsymbol{r}$ is not stored, an attacker must first recover it. Given $\boldsymbol{b}$, $\boldsymbol{r}$ can be recovered from $\boldsymbol{b} \oplus \boldsymbol{u}$. However, since $\boldsymbol{u}$ is not stored, $\boldsymbol{r}$ cannot be recovered. Additionally, the adversary may guess all possible $\boldsymbol{u}$ and $\boldsymbol{r}$ combinations, totaling 2^m possibilities. While this might be feasible for small values of m, in our scheme, $m = 3499$, necessitating 2^{3499} combinations, rendering the attack computationally infeasible. Another method to recover $\boldsymbol{r}$ involves inversely reordering $\boldsymbol{t}$ according to the index vector $\boldsymbol{p}$. However, $\boldsymbol{p}$ is not stored, the adversary cannot use $\boldsymbol{p}$ to recover $\boldsymbol{r}$. Therefore, this is also impossible. Consequently, recovering the feature from a single record is computationally infeasible.

Attack via Record Multiplicity (ARM). In attack via record multiplicity, the adversary knows multiple sets of $\{\boldsymbol{t}, \boldsymbol{b}\}$ generated from the same user in multiple applications, and attempts to recover the original feature $\boldsymbol{f}$ through correlation analysis. As described above, $\boldsymbol{b} = \boldsymbol{u} \oplus \boldsymbol{r}$, thus, an attacker can conduct correlation analysis on multiple $\boldsymbol{b}$ values to reveal $\boldsymbol{r}$ values when the $\boldsymbol{u}$ for different $\boldsymbol{b}$ is identical [22].

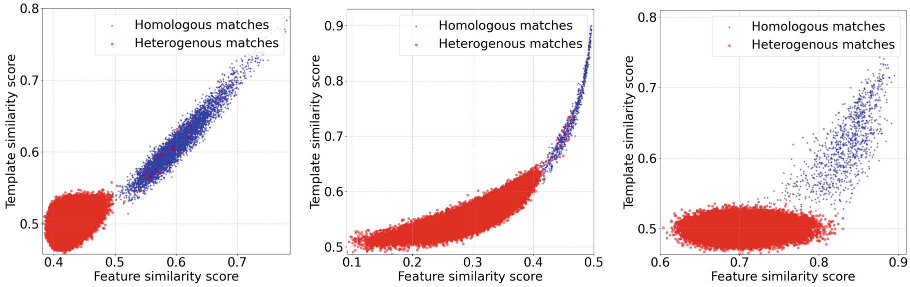
For instance, given $\boldsymbol{f}$ and three distinct $\boldsymbol{r}$, such as $\boldsymbol{r}_A, \boldsymbol{r}_B$ and $\boldsymbol{r}_C$, the generation of different $\boldsymbol{b}$ values is $\boldsymbol{b}_A = \boldsymbol{u}_A \oplus \boldsymbol{r}_A,\ \boldsymbol{b}_B = \boldsymbol{u}_B \oplus \boldsymbol{r}_B,\ \boldsymbol{b}_C = \boldsymbol{u}_C \oplus \boldsymbol{r}_C,$

If the generated hashed code $\boldsymbol{u}$ values remain the same, $\boldsymbol{u}_A = \boldsymbol{u}_B = \boldsymbol{u}_C$, an adversary can cross-XOR the $\boldsymbol{b}$ values as $\boldsymbol{b}_A \oplus \boldsymbol{b}_B = \boldsymbol{r}_A \oplus \boldsymbol{r}_{\mathbf{B}},\ \boldsymbol{b}_A \oplus \boldsymbol{b}_C = \boldsymbol{r}_A \oplus \boldsymbol{r}_{\mathbf{C}},\ \boldsymbol{b}_B \oplus \boldsymbol{b}_C = \boldsymbol{r}_B \oplus \boldsymbol{r}_{\mathbf{C}}$. Afterwards, the attacker can conduct frequency analysis on the relevant $\boldsymbol{r}$ values and quickly recover the possible $\boldsymbol{u}$ and $\boldsymbol{r}$.

However, in practical scenarios, the use of Cosine LSH functions to generate $\boldsymbol{b}$ from $\boldsymbol{f}$ involves a prior randomization process, effectively resisting such attacks. Here, $\boldsymbol{u}$ is a random bitstring generated by $\boldsymbol{f}$ using m Cosine LSH functions. Thus, for the same user, different $\boldsymbol{u}$ values will be formed across different applications, with each bit of $\boldsymbol{u}$ uniformly distributed in $\{0, 1\}$. Consequently, an attacker cannot perform XOR operations to derive $\boldsymbol{u}$ or $\boldsymbol{r}$ from multiple $\boldsymbol{b}$.

Additionally, $\boldsymbol{t}$ is an indexed reordering of $\boldsymbol{r}$. However, given that each $\boldsymbol{r}$ is unique and each bit of $\boldsymbol{r}$ is selected with equal probability, recovering $\boldsymbol{u}$ from multiple sets of $\{\boldsymbol{t}, \boldsymbol{b}\}$ is equally challenging as recovering $\boldsymbol{u}$ from a single $\{\boldsymbol{t}, \boldsymbol{b}\}$ record. Thus, our scheme is irreversible under attacks via record multiplicity.

Similarity-Based Attack (SA). Similarity-based Attack (SA), also known as the Known-Sample attack, has recently been proposed as a threat to cancelable templates [20]. This attack exploits similarity in distance relationships between biometric features before and after transformation to iteratively approximate the feature vectors. To resist this attack, the relationships in similarity scores between heterogenous feature pairs and template pairs should be nonlinear [12]. The relationship is shown in Fig. 6, the red dots indicates heterogenous matches and blue dots indicates homologous matches. The figure shows that in our scheme, the similarity between heterogeneous features and templates primarily exhibits nonlinearity. This suggests that reducing the distance between heterogeneous templates through noise addition or random guessing is challenging. Therefore, the proposed scheme effectively mitigates similarity-based attack.



(a) CASISFace-V5 (b) IITD-E (c) FVC2002 DB3

Fig. 6. The relationship between the similarity scores of features and templates. The x-aix represents Euclidean similarity between features [9], and the y-axis represents Hamming similarity between tamplates.

5.4 Security Analysis

Brute Force Attack (BFA). In security analysis, a brute force attack refers to the possibility of illegal access by using randomly generated templates for authentication. In this scheme, the template is a binary string of length m, resulting in 2^m possible values. With the parameter $m = 3499$, the complexity of successfully guessing the exact template is 2^{3499}, making it infeasible.

False Accept Attack (FAA). In this attack, illegal access can be achieve as long as the similarity scores between the template obtained by the attacker and the enrollment template exceed the threshold τ. Therefore, for $\boldsymbol{t} \in \mathbb{Z}_2^m$ with a matching threshold τ, the expected number of attempts required for a successful attack is $2^{m \times \tau}$. In Fig. 3, the τ for CASIAFace-V5, IITD-E, and FVC2002 DB3 are 0.5292, 0.5981 and 0.5215 respectively. Thus, the attack complexities is at least $2^{3499 \times 0.5215} \approx 2^{1825}$, making it computationally infeasible.

Cipher-Text Only Attack (COA). This attack in symmetric key cryptosystem aims to recover plaintext from ciphertext. In the proposed scheme, a random string $\boldsymbol{r}$ is XORed with a hashed code $\boldsymbol{u}$ to produce $\boldsymbol{b}$, which is stored as ciphertext. Here, $\boldsymbol{r}$ corresponds to the plaintext. Given $\boldsymbol{b} = \boldsymbol{u} \oplus \boldsymbol{r}$, if $\boldsymbol{u}$ is unknown, $\boldsymbol{r}$ remains a uniformly random string. The attack complexity is only related to the length m, which is 2^m, and the attacker cannot learn $\boldsymbol{r}$ from $\boldsymbol{b}$ alone.

Known-Plaintext Attack (KPA). KPA involves recovering the key when both plaintext and ciphertext are known. In our scheme, this equates to determining whether $\boldsymbol{u}$ can be derived from $\boldsymbol{b}$ and $\boldsymbol{r}$. Given $\boldsymbol{b} = \boldsymbol{u} \oplus \boldsymbol{r}$, knowing $\boldsymbol{b}$ and $\boldsymbol{r}$ reveals $\boldsymbol{u}$. However, since $\boldsymbol{r}$ is not stored or accessible to an adversary in our scheme, KPA is not applicable.

6 Conclusion

This paper proposes a one-factor cancelable biometric scheme based on Generated-Index-based Reordering (GI-R) to protect real-valued biometric features. Briefly, the proposed scheme generates an index sequence from a biometric feature and reorders a random string to produce a cancelable template. Therefore, no second factor, such as a token, is required for authentication. It is demonstrated that during template enrollment, each bit of the random string is uniformly selected, ensuring high recognition accuracy and good performance. Experiments are conducted on facial, ear, and fingerprint databases. In addition to performance preservation, the scheme satisfies the requirements of irreversibility, unlinkability and revocability. The security analysis section shows that the proposed scheme effectively resists known attacks.

Acknowledgments. This work is supported by the Strategic Priority Research Program of the Chinese Academy of Sciences, Grant No. XDB0690303. It utilizes the CASIA-FaceV5 dataset from the Chinese Academy of Sciences' Institute of Automation, the IITD-E dataset from Biometrics Research Laboratory at IIT Delhi, and the FVC2002 DB3 dataset. We appreciate the reviewers' positive feedback. We thank Yucheng Ma and Kaini Chen for their valuable discussions on the theorem.

References

1. Casia-facev5. http://biometrics.idealtest.org/
2. Fvc2002. http://bias.csr.unibo.it/fvc2002/
3. Abhishek, N.A.K.N., Akumar, B., Anil, K.J.A.: Biometric template transformation: a security analysis. Proc. SPIE **7541**(3), 175–178 (2010)
4. Charikar, M.: Similarity estimation techniques from rounding algorithms. In: Reif, J.H. (ed.) Proceedings on 34th Annual ACM Symposium on Theory of Computing, 19–21 May 2002, Montréal, Québec, Canada, pp. 380–388. ACM (2002)
5. Dang, T.M., Nguyen, T.D., Hoang, T., Kim, H., Teoh, A.B.J., Choi, D.: Avet: a novel transform function to improve cancellable biometrics security. IEEE Trans. Inf. Forensics Secur. **18**, 758–772 (2022)

6. Deng, J., Guo, J., Xue, N., Zafeiriou, S.: Arcface: additive angular margin loss for deep face recognition. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, 16–20 June 2019, pp. 4690–4699. Computer Vision Foundation/IEEE (2019)
7. Durbet, A., Lafourcade, P., Migdal, D., Thiry-Atighehchi, K., Grollemund, P.M.: Authentication attacks on projection-based cancelable biometric schemes. In: International Conference on Security and Cryptography (2021)
8. Feng, Y., Wang, H., Zhang, D., Li, J., Tao, L.: One-factor cancellable fingerprint template protection based on index self-encoding. J. Database Manag. (JDM) **34**(3), 1–18 (2023)
9. Ghammam, L., Karabina, K., Lacharme, P., Thiry-Atighehchi, K.: A cryptanalysis of two cancelable biometric schemes based on index-of-max hashing. IEEE Trans. Inf. Forensics Secur. **15**, 2869–2880 (2020)
10. Hansley, E.E., Segundo, M.P., Sarkar, S.: Employing fusion of learned and handcrafted features for unconstrained ear recognition. IET Biometrics **7**(3), 215–223 (2018)
11. Indyk, P., Motwani, R.: Approximate nearest neighbors: towards removing the curse of dimensionality. In: Vitter, J.S. (ed.) Proceedings of the Thirtieth Annual ACM Symposium on the Theory of Computing, Dallas, Texas, USA, 23–26 May 1998, pp. 604–613. ACM (1998)
12. Jiang, Y., Shen, P., Zeng, L., Zhu, X., Jiang, D., Chen, C.: Cancelable biometric schemes for euclidean metric and cosine metric. Cybersecurity **6**(1), 4 (2023)
13. Jin, A.T.B., Ling, D.N.C., Goh, A.: Biohashing: two factor authentication featuring fingerprint data and tokenised random number. Pattern Recogn. **37**(11), 2245–2255 (2004)
14. Jin, Z., Hwang, J.Y., Lai, Y.L., Kim, S., Teoh, A.B.J.: Ranking-based locality sensitive hashing-enabled cancelable biometrics: index-of-max hashing. IEEE Trans. Inf. Forensics Secur. **13**(2), 393–407 (2017)
15. Jin, Z., Lim, M., Teoh, A.B.J., Goi, B., Tay, Y.H.: Generating fixed-length representation from minutiae using kernel methods for fingerprint authentication. IEEE Trans. Syst. Man Cybern. Syst. **46**(10), 1415–1428 (2016)
16. Kim, J., Teoh, A.B.J.: One-factor cancellable biometrics based on indexing-first-order hashing for fingerprint authentication. In: 2018 24th International Conference on Pattern Recognition (ICPR), pp. 3108–3113. IEEE (2018)
17. Kumar, A., Wu, C.: Automated human identification using ear imaging. Pattern Recogn. **45**(3), 956–968 (2012)
18. Lacharme, P.: Analysis of the iriscodes bioencoding scheme. Int. J. Comput. Sci. Softw. Eng. (IJCSSE 2012) **6**(5), 315–321 (2012)
19. Lai, Y.L., et al.: Cancellable iris template generation based on indexing-first-one hashing. Pattern Recogn. **64**, 105–117 (2017)
20. Lai, Y., Jin, Z., Wong, K., Tistarelli, M.: Efficient known-sample attack for distance-preserving hashing biometric template protection schemes. IEEE Trans. Inf. Forensics Secur. **16**, 3170–3185 (2021)
21. Lee, M.J., Jin, Z., Li, M., Chen, D.B.W.: Mixing binary face and fingerprint based on extended feature vector (EFV) hashing. In: 2019 International Symposium on Intelligent Signal Processing and Communication Systems (ISPACS), pp. 1–2. IEEE (2019)
22. Lee, M.J., Jin, Z., Teoh, A.B.J.: One-factor cancellable scheme for fingerprint template protection: extended feature vector (EFV) hashing. In: 2018 IEEE International Workshop on Information Forensics and Security (WIFS), pp. 1–7. IEEE (2018)

23. Lee, M.J., Teoh, A.B.J., Uhl, A., Liang, S.N., Jin, Z.: A tokenless cancellable scheme for multimodal biometric systems. Comput. Secur. **108**, 102350 (2021)
24. Lee, Y., Chung, Y., Moon, K.: Inverse operation and preimage attack on biohashing. In: 2009 IEEE Workshop on Computational Intelligence in Biometrics: Theory, Algorithms, and Applications, pp. 92–97. IEEE (2009)
25. Li, H., Wang, X.: One factor cancellable fingerprint scheme based on novel minimum hash signature and secure extended feature vector. Multimedia Tools Appl. **81**(9), 13087–13113 (2022). https://doi.org/10.1007/s11042-022-12424-y
26. Manisha, Kumar, N.: Cancelable biometrics: a comprehensive survey. Artif. Intell. Rev. **53**(5), 3403–3446 (2020)
27. Ouda, O., Tsumura, N., Nakaguchi, T.: Tokenless cancelable biometrics scheme for protecting iris codes. In: 2010 20th International Conference on Pattern Recognition, pp. 882–885. IEEE (2010)
28. Pillai, J.K., Patel, V.M., Chellappa, R., Ratha, N.K.: Sectored random projections for cancelable iris biometrics. In: 2010 IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 1838–1841. IEEE (2010)
29. Ratha, N.K., Connell, J.H., Bolle, R.M.: Enhancing security and privacy in biometrics-based authentication systems. IBM Syst. J. **40**(3), 614–634 (2001)
30. Sadhya, D., Akhtar, Z., Dasgupta, D.: A locality sensitive hashing based approach for generating cancelable fingerprints templates. In: 2019 IEEE 10th International Conference on Biometrics Theory, Applications and Systems (BTAS), pp. 1–9. IEEE (2019)
31. Sadhya, D., Raman, B.: Generation of cancelable iris templates via randomized bit sampling. IEEE Trans. Inf. Forensics Secur. **14**(11), 2972–2986 (2019)
32. Wang, X., Li, H.: One-factor cancellable palmprint recognition scheme based on oiom and minimum signature hash. IEEE Access **7**, 131338–131354 (2019)

Multi-teacher Invariance Distillation for Domain-Generalized Action Recognition

Jongmin Shin[1], Abhishek Maiti[2], Yuliang Zou[3], and Jinwoo Choi[1](✉)

[1] Kyung Hee University, Yongin-si, Gyeonggi-do 17104, Republic of Korea
jinwoochoi@khu.ac.kr
[2] IIIT Delhi, New Delhi 110020, India
[3] Virginia Tech, Blacksburg, VA 24061, USA

Abstract. In this work, we tackle the problem of domain-generalized action recognition, i.e. we train a model on a source domain and then test the model on other unseen target domains with different data distributions. Generalizing across different domains often requires distinct representational invariances and variances, which makes domain generalization even more challenging. However, existing methods overlook the nuanced requirements of representational invariances/variances across different domains. To this end, we propose Multi-teacher Invariance Distillation for domain-generalized Action Recognition (MIDAR), a method to learn multiple representational invariances/variances tailored to the unique characteristics of diverse domains. MIDAR comprises two key learning stages. First, we learn multiple teacher models to specialize in distinct representational invariances/variances. Then, we distill the knowledge of teachers to a student model through the adaptive reweighting (ARW) layer, which determines the ratio of supervision from different teachers. We validate the proposed method on public benchmarks. The proposed method shows favorable performance compared to the existing methods across multiple domains on public benchmarks.

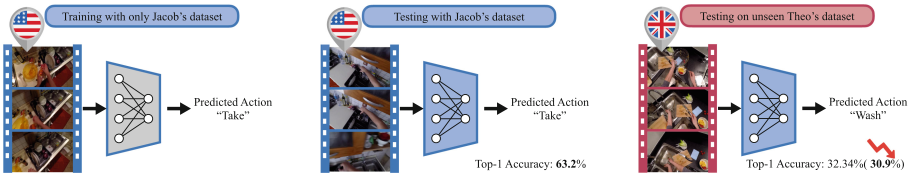
Keywords: Action Recognition · Domain Generalization · Knowledge Distillation · Self-Supervised Learning · Invariance

1 Introduction

The rapid progress in action recognition [6,15,16,26,41,52] has significantly improved the ability of video models to understand human actions in videos. Despite the great progress, most action recognition models often suffer from performance degradation on the test datasets with different distributions from the training dataset [9–11,33]. This performance drop is evident in domain generalization [47], highlighting the vulnerability of action recognition models to distribution shifts. As shown in Fig. 1 (a) and (b), training and testing in the same dataset, e.g. Jacob's kitchen dataset, allows the model to correctly recognize the action '*Take*'. However, as depicted in Fig. 1 (c), testing the

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/978-3-031-78110-0_8.

A. Antonacopoulos et al. (Eds.): ICPR 2024, LNCS 15329, pp. 116–132, 2025.
https://doi.org/10.1007/978-3-031-78110-0_8


Training with only Jacob's dataset
Predicted Action "Take"
Testing with Jacob's dataset
Predicted Action "Take"
Top-1 Accuracy: 63.2%
Testing on unseen Theo's dataset
Predicted Action "Wash"
Top-1 Accuracy: 32.34%(30.9%)

(a) Model training (b) In-domain testing (c) Out-of-domain testing

Fig. 1. (Single-source) domain generalization. (a) We train a video recognition model on a source domain (e.g., Jacob's Kitchen); (b) When we test the model on the same data distribution, the model performs reasonably well; (c) However, when the model is evaluated on an unseen target domain (e.g., Theo's Kitchen), the performance drops significantly due to domain shifts.

model on a dataset with a distribution shift from the training data, e.g. trained on Jacob's kitchen dataset and test on Theo's kitchen dataset, significantly degrades the model performance from 63.2% to 30.9%. The model fails to recognize the action '*Take*' and misclassifies it as '*Wash*'. A desired model would not suffer from this performance drop across domains.

We hypothesize that we can enhance the generalization performance of a model by learning multiple representational invariances/variances. We empirically find that the beneficial invariances/variances depend on the source and target distributions. In Table 1, we show domain generalization performance of a few models: i) a baseline TSM [26] that does not explicitly learn any representational invariances, ii) a color-invariant TSM, iii) a temporal-invariant TSM, iv) an color&temporal-invariant TSM, all evaluated on the EPIC-KITCHENS dataset [12]. Please refer to Sect. 3.1 for detailed information on model training procedures. We find that the color and temporal-invariant model outperforms the baseline, whereas the color&temporal-invariant model underperforms the baseline. The results indicate that the effectiveness of specific invariances/variances depends on the source and target distributions. We could expect improved generalization performance if we can appropriately learn to incorporate multiple invariances/variances.

Multi-source domain generalization [2,24,25,40] might be a solution to learn multiple invariances. However, it is impractical for video action recognition as collecting and labeling multiple video action recognition datasets is labor-intensive and costly. Single-source domain generalization methods [5,8,42,43,47,53] could learn representational invariances in image recognition. However, we empirically find these methods struggle with the temporal dimension critical for video data. RADA [47] learns invariances by adversarial perturbations. They perturb

Table 1. Baseline Domain Generalized Action Recognition Performance. We show the domain generalization accuracy of models with distinct representational invariances and a model naively learned multiple invariances. We use the TSM model with a ResNet-50 backbone.

Method	Average Accuracy
Baseline Model	37.07 ± 3.39
Color Invariant Model	37.83 ± 3.65
Temporal Invariant Model	38.36 ± 2.73
Color&Temporal Invariant Model	35.34 ± 5.38

the data distribution of the source domain to cover the unseen target domain. However, they do not learn diverse invariances e.g. temporal and order variance/invariance, which may be beneficial in some domains. A naive approach for learning multiple invariances could be training a model with multiple tasks, each responsible for a specific type of invariance. However, we empirically find this approach results in inferior performance even compared to the baseline without any invariances in Table 1. Models with only a single type of invariance, e.g. color invariance, show improved domain generalization performance (38.36% *vs.* 37.07%). However, a model naively trained with both color and temporal invariance learning heads underperforms compared to the baseline (35.34% *vs.* 37.07%). The observations underscore the importance of a nuanced approach to learning multiple representational invariances and variances to achieve robust performance across diverse domain generalization scenarios.

In this work, we introduce Multi-teacher Invariance Distillation for domain-generalized Action Recognition (MIDAR) to address the challenge of learning multiple invariance/variance. Our approach involves two stages. In the first stage, we train multiple teacher models, each specializing in a different representational invariance or variance. In the next stage, we distill the knowledge from multiple teachers into a student model. MIDAR adaptively reweighs the supervision from multiple teachers, allowing the student model to learn distinct representational invariance/variance. We validate the effectiveness of the proposed method on public benchmarks. MIDAR shows favorable performance compared to the existing methods.

To summarize, we make the following contributions.

- We introduce MIDAR, a new training method addressing the challenge of learning multiple representational invariances/variances for domain-generalized action recognition.
- We introduce the Adaptive Reweighting layer to adjust the contribution of multiple teachers, allowing the student model to leverage the diverse representational invariance/variance of each teacher.
- We conduct extensive experiments on the Epic-Kitchens benchmark to validate MIDAR. Our findings indicate that MIDAR's approach to learning diverse representational invariance/variance outperforms current SOTA methods like RADA, which rely on adversarial perturbation.

2 Related Work

2.1 Video Action Recognition

2D CNNs [26,41,52], 3D CNNs [6,15,38], and two-stream CNNs [16,34] are popular techniques to recognize human actions from videos. More recently, Transformer-based methods have shown great performance [3,4,14,18,31,44,46]. Despite the great recent advances in action recognition, we find that state-of-the-art action recognition methods still suffer from cross-domain generalization: a model trained on one domain shows poor performance on other domains with different data distributions. In this work, we tackle the domain-generalized action recognition task to address the challenge.

2.2 Domain Generalization

Recently, domain generalization has drawn significant attention from the community since training and test data usually have different distributions in practice. Broadly, there are two principal categories of approaches in the domain generalization literature: i) feature-based domain generalization, and ii) data-based domain generalization. Feature-based domain generalization methods [2,5,24,25,42] aim to learn domain-invariant representations to enhance the generalization performance of models. On the other hand, data-based domain generalization methods [21,39,40] augments training data to generate adversarial samples and synthetic data with different styles and scenes that bridge the gap between source and target domains. These works have shown great progress in domain-generalized image recognition. However, domain generalization for video recognition is still under-explored. To the best of our knowledge, there is only one work on domain-generalized action recognition: Robust Adversarial Domain Augmentation (RADA) [47]. RADA learns domain invariant video representation by training on the perturbed data and adversarial examples. Our work is on domain-generalized action recognition as well. In contrast to RADA, the proposed method learns the nuanced requirements of the representational invariances across different domains, thus offering a novel approach to this challenging problem.

2.3 Knowledge Distillation

Knowledge distillation is a popular technique to transfer knowledge from one model to another model. We can categorize knowledge distillation into three groups: i) response-based, ii) intermediate, iii) relation-based, and iv) multi-teacher knowledge distillation. Response-based knowledge distillation methods [19,51,54] encourage the student model to mimic the output of the teacher. In intermediate knowledge distillation [1,29], the student model aims to learn the same feature representation as the feature representation of the teacher. In Relation-based knowledge distillation [30], the student model mimics the relative distance and angle between data points in the feature space of the teacher model. In Multi-teacher knowledge distillation [27,48,49], a student model learns from the combined knowledge of multiple teacher models, leveraging diverse representations. In this work, we leverage knowledge distillation techniques to address the challenge of domain generalization. We learn a student model using multiple teachers, each of which specializes in distinct representational invariances. We dynamically adjust the contribution of different teachers by learning an adaptive re-weighting layer.

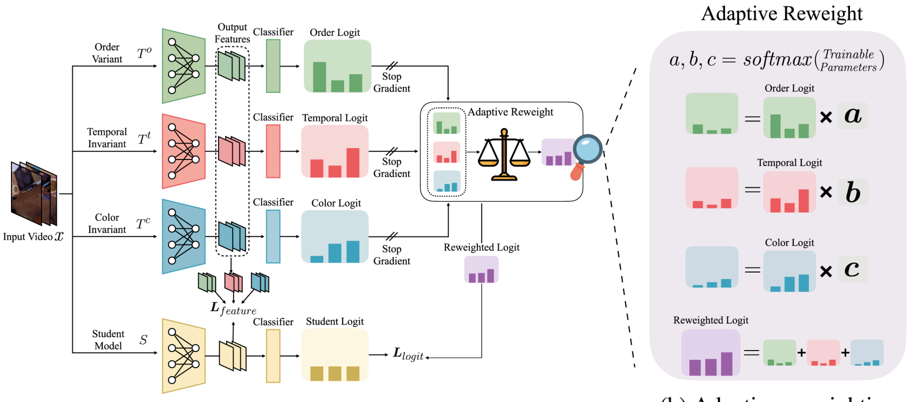
3 Method

We propose Multi-teacher Invariance Distillation for domain-generalized Action Recognition (MIDAR). As shown in Fig. 2, we employ multiple teachers each with expertise in distinct representational invariance/variance. Our objective is to distill a broad spectrum of invariances/variances, including order variance, temporal invariance, and color invariance, into a student model. This knowledge distillation process encompasses both the feature representations and the output logits of these teacher models. We propose an adaptive reweighting layer to dynamically adjust the contribution from each teacher based on the data. In the following subsections, we provide detailed descriptions of each component of MIDAR. We describe the training process of teacher models in Sect. 3.1. Then we illustrate the proposed multi-teacher distillation framework in Sect. 3.2. Finally, we describe the proposed adaptive reweighting method in Sect. 3.3.

3.1 Training Teacher Models

Color Invariant Teacher. Color invariance is desirable in many action recognition scenarios. For example, a model should be able to correctly recognize the 'playing tennis' action regardless of whether the tennis court is green grass or brown mud. To learn color invariance, we employ color jittering augmentation during the color invariant teacher training process. Given an input video, we randomly jitter the brightness, contrast, saturation, and hue of each frame. Following prior works [17,32,55], we employ a temporally coherent color jitter augmentation, i.e. we use the same color jittering across all the frames within an input video.

We employ supervised contrastive learning (SCL) [23] for color invariant teacher training. We empirically find that SCL is beneficial for color invariance learning, compared to using the cross-entropy loss. In the SCL framework, we define any pair of videos from the same action class as a positive pair, regardless of color augmentation. We define any pairs from different action classes as negative pairs. We define the SCL loss for learning color invariance as follows:



(a) Multi-teacher distillation

(b) Adaptive reweighting layer

Fig. 2. Overview. (a) We use a multi-teacher distillation framework to distill multiple representational (in)variances into a student model. Both features and the logits are distilled from each teacher to the student model. (b) For logit distillation, we propose an adaptive reweighting layer to adjust the impacts of each teacher. Specifically, we assign one learnable parameter for each teacher so that the distillation strength of each teacher is dynamically adjusted during training.

$$L_{\text{SCL}} = \sum_{i \in B} \frac{-1}{|P(i)|} \sum_{p \in P(i)} \log \frac{\exp(\mathbf{z}_i \cdot \mathbf{z}_p / \tau)}{\sum_{a \in A(i)} \exp(\mathbf{z}_i \cdot \mathbf{z}_a / \tau)}, \tag{1}$$

where B denotes the set of all input data within a minibatch while each i-th instance is an anchor. $A(i)$ denotes the set of all input data within a mini-batch except the i-th instance, i.e. $A(i) \equiv B \backslash \{i\}$. The set of all positive pairs $P(i)$ contains samples with identical class labels to i-th instance, including color-augmented samples. In (1), we scale the similarity between the anchor embedding $\mathbf{z}_x$ and any positive sample embedding $\mathbf{z}_p$ by the temperature hyperparameter τ.

In SCL, a model learns to align positive pairs, each consisting of different augmentations. As a result, a teacher model trained with SCL has specialized expertise, i.e. color invariance in our case, that could be generalizable across different domains [37].

Temporal Invariant Teacher. Unlike image data, video data has an additional temporal dimension. The same human action might have different speeds, durations, or temporal patterns across different domains. Consequently, to robustly recognize human actions in various domains, we desire a temporal invariant model [13,17,35,55]. To learn temporal invariant representations to a teacher model, we employ three temporal augmentations [55] that have shown significant performance improvement: T-Half, T-Drop, and T-Reverse. For example, let us assume we have 4 frames with indices $[1, 2, 3, 4]$. Then the T-Half augmentation repeats the first or the second half of the video only: e.g. $[1, 2, 3, 4] \rightarrow [1, 2, 1, 2]$ or $[1, 2, 3, 4] \rightarrow [3, 4, 3, 4]$. The T-Half augmentation encourages the model to be robust to the partial temporal occlusion. The T-Drop augmentation drops random frames in the video, substituting them with the previous frame: e.g. $[1, 2, 3, 4] \rightarrow [2, 2, 4, 4]$. The T-Drop augmentation encourages the model to be invariant to the speed of the action. The T-Reverse augmentation inverts the order of the video frames, e.g. $[1, 2, 3, 4] \rightarrow [4, 3, 2, 1]$. Following the prior work [55], we randomly select one temporal augmentation for augmenting each video. We employ (1), supervised contrastive learning with these temporal augmentations [55]. We empirically find that using the SCL loss is beneficial for temporal invariance learning compared to using the standard supervised training with the cross-entropy loss for the prediction.

Order Variant Teacher. To distinguish fine-grained actions with subtle differences, e.g. opening and closing a door, a model needs to be sensitive to the temporal order of events. To encourage a model to be sensitive to the order of temporal events, we employ a self-supervised task: video clip order prediction [45]. In this task, we shuffle the clips sampled from an input video. Then we input the shuffled clips into a model. The model should predict the correct chronological order of the clips. Predicting the correct temporal sequence of video clips encourages the model to specialize in order *variance*. Through this task, a model better understands temporal relationships and dependencies between different temporal segments of the actions. Order variance is beneficial for domain generalized action recognition since the order of the action often does not change across people or locations. Furthermore, order variance is desirable as learning order variant representation is learning action representations robust to scene distribution shift across domains [11].

To learn order variance, we define the order variance (OV) loss as follows:

$$L_{\text{OV}} = -\sum_{i=1}^{C!} y_i \log f_o(\psi). \tag{2}$$

Here, the model takes the concatenated input $\psi = (\phi_1, ..., \phi_C)$, where each ϕ_i is a feature vector of i-th clip in the input video. The model predicts a probability distribution across $C!$ possible temporal orders of the input clips, where C denotes the number of clips in an input video. y_i is the i-th element of $\mathbf{y}$, while $\mathbf{y}$ is the ground truth one-hot vector of length $C!$ with the correct clip order of the input video.

3.2 Distilling Invariances from Multiple Teachers

In Table 1, we observe that the naive learning of multiple representational invariances degrades the domain generalized action recognition performance (i.e. 35.34 ± 5.38 vs. 37.07 ± 3.39). To address this challenge, we propose Multi-teacher Invariance Distillation for domain-generalized Action Recognition (MIDAR). MIDAR has a multi-teacher knowledge distillation architecture [20,22,27] comprising teacher models with expertise in order variance (Ω_O), temporal invariance (Ω_T), and color invariance (Ω_C). As depicted in Fig. 2 (a), each teacher contributes distinct expertise to the learning process.

Both the student model, π, and the teacher models, Ω_O, Ω_T, and Ω_C, take the same input RGB video, $I \in \mathbb{R}^{M \times H \times W \times C}$, where each video contains M frames with the height of H pixels, width of W pixels, and C channels. We employ the feature-space distillation loss, L_{feature}. We compute L_{feature} as the mean squared error between the feature vectors of the student and teacher models as follows:

$$L_{\text{feature}} = \sum_{t \in \{O,T,C\}} \left(\frac{1}{n} \sum_{i=1}^{n} (\Omega_t(i) - \pi(i))^2 \right). \tag{3}$$

Here, t is the teacher model index and $\Omega(i)$ and $\pi(i)$ denote the i-th feature of the teacher and student model, respectively. By using the L_{feature}, we effectively guide the student model to mimic the expertise of the teachers.

Moreover, we employ the Kullback-Leibler (KL) Divergence loss, L_{KL}, in MIDAR for the output-space distillation as follows:

$$L_{\text{KL}} = \sum_{k=1}^{K} P_\Omega(k) \log \left(\frac{P_\Omega(k)}{Q_\pi(k)} \right). \tag{4}$$

Here P_Ω denotes the output probability of the adaptive reweighting (ARW) layer and k is the action category index for K action categories. Q_π is the output probability of the linear action classifier for the student model. The output-space distillation encourages the student model to mimic the prediction of the teacher model.

3.3 Learning to Re-weight Multiple Teachers

We introduce an adaptive reweighting (ARW) layer in MIDAR to reflect the nuanced influence of multiple teacher models for learning a student model. The ARW layer takes the softmax probability of each teacher Ω_O, Ω_T, and Ω_T and outputs the single softmax probability vector P_Ω. We define adaptive reweighting operation as follows:

$$P_\Omega = \sum_{t \in \{O,T,C\}} \frac{\exp(\alpha_t)}{\sum_{i \in O,T,C} \exp(\alpha_i)} f_t(\Omega_t). \tag{5}$$

Here, f_t is a linear action classifier for the teacher t. α_i's are learnable parameters for the adaptive reweighting. We set the same number of parameters the same as the number of teachers. By (5), we get the final reweighted probability, P_Ω, that aggregates the nuanced contributions of all the teacher models, as illustrated in Fig. 2 (b).

During training, the parameters α_i's are continuously updated, leading to dynamic adjustments of the contribution of each teacher: more effective teachers get higher weights and less effective ones get lower weights. The learnable parameter α_i dynamically adjusts the student model's focus on multiple invariances and variances. The balance is crucial for enhancing the student model performance, as it allows for a more nuanced understanding that could be beneficial in multiple domains. The proposed ARW layer enables the student model to effectively extract the diverse representational invariances/variances of the teacher models.

We define the total loss function of MIDAR as follows:

$$L = L_{\text{CE}} + L_{\text{feature}} + L_{\text{KL}}. \tag{6}$$

The total loss function consists of three components. First, L_{CE} is the standard cross-entropy loss to learn action categories. L_{feature} aligns feature representations of the student model with the feature representations of the teacher models. L_{KL} guides the student model to mimic the adaptively re-weighted predictions of the teacher models.

4 Experimental Results

4.1 Experimental Setup

Dataset. To evaluate the effectiveness of MIDAR, we use the EPIC-KITCHENS-55 dataset [12]. EPIC-KITCHENS-55 is a large-scale egocentric action recognition dataset consisting of multiple domains. We use the subset for evaluating domain generalization methods, following the experimental protocol in a prior work [28]. The subset comprises three domains, D1, D2, and D3, which results in six domain generalization settings: D1→D2, D1→D3, D2→D1, D2→D3, D3→D1, and D3→D2. The subset consists of 8 action classes across all the domains: put, take, open, close, wash, cut, mix, and pour. Each domain has different actors and kitchen environments but the same action categories. The subset consists of $10,094$ videos in total.

Evaluation Metric. We evaluate the effectiveness of Multi-teacher Invariance Distillation for domain-generalized Action Recognition (MIDAR) by adopting the standard evaluation protocols across benchmarks [28]. For the Epic-Kitchens benchmark, in our protocol [28], we select the model that demonstrates the highest in-domain performance and evaluate the cross-domain performance of the model. We measure the model's performance using the averaged Top-1 accuracy and the standard deviation across six different cross-domain generalization settings.

Implementation Details. Here, we provide details of our training setup and implementation. For additional information, please refer to the supplementary materials. *Base setting.* We employ Temporal Shift Module (TSM) [26] with a ResNet-50 backbone as the base model, unless we specify another model. From each video, we sample 8 frames to construct an input clip. The initial learning rate is 0.0075. We train models for 150 epochs.

Teacher Model Training. For color and temporal invariant teacher models, we build the models upon the SimSiam [7] architecture. We use the supervised contrastive loss Eq. (1) as a loss function to train the color-invariant and temporal-invariant teacher models, with the temperature τ set to 0.3. For the order variant teacher, we implement the video clip order prediction (VCOP) [45] pre-text task, processing 3 clips of 8 frames each, with an inter-clip interval of 8 frames. We train the model for 800 epochs. We attach a linear classifier on top of the backbone. Then we train the model end-to-end just like other teacher models.

Student Model Training. When training the student model, we freeze the weights of all the teacher models. The learning rate is set to 0.005. For the adaptive reweighting layer, each trainable parameter α_t is initially set to an equal value of 1. This initialization strategy ensures that, before updating the trainable parameters, each value post-softmax normalization approximates 0.3333, thereby providing a fair starting point.

Please see the supplementary materials for details on the model training and inference.

Baseline. To establish a baseline for domain generalization performance, we train a TSM with a ResNet-50 backbone on one domain of the benchmark dataset. Subsequently, we evaluate the trained model on another domain of the dataset. We repeat the same process for all six settings in the EPIC-KITCHENS dataset. During training, we do not apply any learning technique that encourages domain-invariant representations. To establish a baseline for domain generalization performance, we train a TSM with a ResNet-50 backbone on one

Table 2. Individual Invariant/Variant Model Performance. We show the domain generalization performance of individual invariant/variant models. Every model is equipped with the TSM with a ResNet-50 backbone.

Method	Top-1 Accuracy
Baseline	37.07 ± 3.39
Color Invariant Model	37.83 ± 3.65
Temporal Invariant Model	**38.36 ± 2.73**
Order Variant Model	37.38 ± 3.54

domain of the EPIC-KITCHENS dataset [28]. Subsequently, we evaluate the trained model on another domain of the EPIC-KITCHENS dataset. We perform the same process for all six settings in the EPIC-KITCHENS dataset. During training, we do not apply any learning technique that encourages domain-invariant representation learning.

4.2 Individual Invariance/Variance Model Performance

We first study the effectiveness of each model with distinct representational invariances/variances by comparing the domain generalization performance of each model with the baseline performance. As shown in Table 2 each representational invariant/variant model outperforms the baseline. The temporal invariant model shows the most improvement of 1.29 points and the color invariant model shows an improvement of 0.76 points compared to the baseline. The order variant model achieves a marginal improvement of 0.31 points compared to the baseline. These results indicate that incorporating individual representational invariances/variances could improve the domain generalization performance, but the improvement is not very significant. As shown in Table 1, naively learning multiple invariances, e.g. learning the temporal and the color invariant representations simultaneously, results in inferior performance compared to the baseline without learning any invariances: 35.34% *vs.* 37.07%. Therefore, we need a nuanced approach, such as MIDAR, to learn multiple representational invariances/variances to achieve superior domain generalization performance.

Table 3. Effect of Distillation in Domain Generalization. We compare the performance of the logit-space distillation, the feature-space distillation, and both the logit-space and faeture-space distillation. We employ the temporal invariant model as a teacher in this experiment.

Method	Logit	Feature	Top-1 Accuracy
Baseline	–	–	37.07 ± 3.39
Temporal Invariant Teacher	–	–	38.36 ± 2.73
Student	✓		38.78 ± 2.85
		✓	35.28 ± 4.13
	✓	✓	**38.93 ± 3.61**

Table 4. Effect of Multi-Teacher Distillation. We compare the domain generalization performance of students learned from different combinations of teachers. Properly using all three teachers shows the best domain generalization performance.

Method	Color	Temporal	Order	Top-1 Accuracy
Baseline	–	–	–	37.07 ± 3.39
Single Teacher Distillation		✓		38.93 ± 3.61
Two-Teacher Distillation	✓	✓		38.64 ± 2.48
	✓		✓	37.20 ± 5.30
		✓	✓	40.03 ± 3.29
Three-Teacher Distillation	✓	✓	✓	**41.12 ± 2.61**

Table 5. Ablation study. We conduct experiments with different distillation methods to validate the effect of each distillation strategy, logit, feature, and multi-teacher distillation.

(a) How to aggregate multiple teacher outputs?

Method	Top-1 Accuracy
Baseline	37.07 ± 3.39
Correct Teacher	38.55 ± 1.74
Most Confident Teacher	38.99 ± 3.76
Lowest Cross-Entropy Teacher	38.89 ± 2.81
Average of Teachers	38.70 ± 3.74
Adaptive Reweighting (Ours)	41.12 ± 2.61

(b) Which loss to distill features?

Method	Top-1 Accuracy
Baseline	37.07 ± 3.39
CORAL Loss [36]	40.12 ± 3.47
Huber Loss	39.02 ± 3.87
MSE Loss	41.12 ± 2.61

(c) Multi-teacher distillation method

Method	Top-1 Accuracy
Baseline	37.07 ± 3.39
KD [19]	38.45 ± 3.86
FiTNet [1]	37.38 ± 4.34
Average [48]	38.68 ± 3.70
Ours	41.12 ± 2.61

4.3 Distillation for Learning Multiple Invariances/variances

Is Distillation Beneficial in Domain Generalization? We empirically find that distillation is beneficial in domain generalization. In Table 3, compared to the temporal invariant teacher model, a student model learned by the logit-space distillation shows an improved performance of 38.78%. A student model learned by the feature-space distillation shows inferior performance compared to the teacher. However, a student model learned by both logit and feature-space distillation shows the best performance of 38.93%. The results demonstrate that distillation is beneficial in domain generalization even when we have a single teacher only. In the remaining experiments, we distill both features and logits.

Is Multi-Teacher Distillation Beneficial for Learning Multiple Invariances/variances? We investigate the effect of multiple teachers in Table 4. We can observe a trend: as the number of teachers increases, the student model demonstrates improved performance. Specifically, the student model, which learns from both the temporal-invariant and order-variant teachers, achieves an accuracy of 40.03%, surpassing the single teacher distillation with an accuracy of 38.93%. Furthermore, when the student model learns from the knowledge distillation of the color invariant, the temporal invariant, and the order variant teachers simultaneously, the student model achieves the best accuracy of 41.12%. The results underscore the significance of leveraging multiple teachers to enrich the knowledge of the student model and subsequently enhance the domain generalization.

4.4 Ablation Study

We conduct ablation experiments to explore the various design choices of the multi-teacher distillation strategy to improve the domain generalization performance. Here, we conduct all experiments with multi-teacher distillation that encompasses all invariant and variant teacher models.

How to Aggregate the Output of Multiple Teachers? Since we have multiple teachers, how to aggregate the output of multiple teachers is an important design choice. In Table 5 (a), we compare five logit-space distillation methods. i) Correct Teacher: we select the correctly predicted teachers for the distillation. We average the prediction vectors if multiple teachers agree, and we discard the sample if all predictions are incorrect. ii) Most Confident Teacher: we select the teacher with the highest softmax probability among all the teachers for the distillation. iii) Lowest Cross-Entropy Teacher: we choose the prediction from the teacher with the minimum cross-entropy loss for the distillation. iv) Average of Teachers: we average the predictions of all the teachers for the distillation. v) Adaptive reweighting (ours): we dynamically adjust the contribution of each teacher by Eq. 5. The results demonstrate that the adaptive reweighting outperforms the other compared methods, achieving 41.12% which is 2.13 points higher than the second-best method, Most Confident Teacher. The results suggest that our adaptive reweighting is more effective in leveraging multiple teachers to improve domain generalization.

Which Loss for Feature-Space Distillation? Here, we compare three loss functions for the feature-space distillation in MIDAR. i) The CORAL (CORelation ALignment) loss [36]: we align the second-order statistics of feature distributions by minimizing the difference in their covariance matrices. The CORAL loss is typically applied for domain adaptation. We employ the CORAL loss to align student features with the teacher features to tackle the problem of domain generalization. ii) Huber loss is a hybrid loss function that is a combination of both Mean Squared Error (MSE) and Mean Absolute Error (MAE). Huber loss aims to mitigate the influence of outliers during distillation. Also, huber loss offers a balance between sensitivity to data variance and robustness to outliers. iii) Mean Squared Error (MSE) loss: a loss function that minimizes the average of the squares of the errors. As shown in Table 5 (b), employing CORAL loss outperforms Huber loss with a margin of 1.1 points (40.12% vs. 39.02%). However, using MSE loss outperforms CORAL loss with a margin of 1.0 points (41.12% vs. 40.12%). The results suggest that the MSE loss is more effective for feature-space distillation.

Comparing Multi-teacher Distillation Methods. We conduct a comparative analysis of MIDAR against established distillation techniques. We replace the proposed multi-teacher distillation method with the following methods and compare the domain generalization performance. i) KD [19], which distills the average predictions from multiple teachers, ii) FitNet [1], which distills their average features, and iii) Average [48], which distills both averaged features and predictions. Table 5(c) shows that MIDAR achieves the best accuracy of 41.12%. Compared to MIDAR, FitNet shows a 3.74-point drop (41.12% *vs.* 37.38%) and KD shows a 2.67-point drop (41.12% *vs.* 38.45%). Average shows a 2.44-point drop (41.12% *vs.* 38.68%). The results showcase the effectiveness of our multi-teacher distillation approach in enhancing domain generalization.

4.5 Comparison with Existing Domain Generalization Methods

We compare the domain generalization performance of MIDAR with existing single-source domain generalization methods in Table 6. Please see the supplementary materials for details of the results. We compare MIDAR with four image-based methods extended to the video domain. i) Mixup [50]: we blend each video with a randomly chosen video in the batch and set the mixup ratio as 0.2. ii) Mixstyle [53]: we integrate a Mixstyle module into the ResNet backbone of TSM. Mixstyle mixes the statistics, i.e. mean and standard deviation, of feature maps from different instances during the training process. Mixstyle incorporates a new style in the feature space and encourages the model to learn domain generalizable features. iii) JiGen [5] recognizes action and simultaneously solves jigsaw puzzles to understand spatial correlations. Solving jigsaw puzzles acts as a regularization for the classification task. The shared feature embedding between the classification and the jigsaw puzzle tasks allows the model to generalize across domains. iv) EISNet [42] enhances generalization performance by multi-task learning from both extrinsic and intrinsic supervisions. EISNet employs momentum metric learning for domain-invariant yet class-specific features and solves jigsaw puzzles. For JiGen and EISNet baselines, we apply consistent augmentation across all frames in a video clip to maintain temporal coherence. Additionally, we compare MIDAR with RADA [47][1], the state-of-the-art video-based domain generalization method.

Table 6. Comparison with the state of the arts on EPIC-Kitchens. We compare the domain generalization performance (top-1 accuracy) of our model with several image-based methods (Mixup [50], Mistyle [53], JigSen [5], EISNet [42]) and a recently proposed video-based method (RADA [47]).

Method	Backbone	Average Accuracy
Baseline	TSM	37.07 ± 3.39
Mixup [50]	TSM	37.54 ± 4.69
Mixtyle [53]	TSM	36.88 ± 5.18
JiGen [5]	TSM	38.59 ± 6.14
EISNet [42]	TSM	37.52 ± 1.31
RADA [47]	APN [47]	40.52 ± 3.23
Ours	TSM	**41.12 ± 2.61**

All the compared methods employ the TSM [26] as a backbone except RADA. RADA [47] is equipped with the Adversarial Pyramid Network (APN) backbone. We select the learning rate with the highest performance for each method. We use the learning rates of 0.01, 0.001, 0.005, and 0.0075 for Mixup, Mixstyle, JiGen, and EISNet respectively. For RADA we use the learning rate of 0.001. As shown in

Table 7. Effect of using different backbones: ResNet-50 *vs.* ResNet-101 on the EPIC-Kitchens dataset. We compare the domain generalization performance (top-1 accuracy) of our model with the video-based method(RADA [47]).

Method	Model	Backbone	Average Accuracy
RADA [47]	APN [47]	ResNet-50	40.52 ± 3.23
Ours	TSM	ResNet-50	**41.12 ± 2.61**
RADA [47]	APN [47]	ResNet-101	43.08 ± 4.27
Ours	TSM	ResNet-101	**43.54 ± 5.59**

[1] The TSM backbone employed by MIDAR has 4.8 million *fewer* parameters than the APN used by RADA.

Table 6, JiGen outperforms other image-based methods with an average accuracy of 38.59% exceeding Mixstyle by 1.71 points, Mixup by 1.05 points and EISNet by 1.07 points.

However, JiGen shows inferior performance compared to the video-based method RADA by 1.93 points. MIDAR surpasses RADA by 0.60 points and Jigen by 2.53 points, resulting in the best accuracy as well as relatively lower standard deviation. The results indicate that MIDAR shows favorable performance across various domain shifts and demonstrates the effectiveness in various domain generalization scenarios.

For a fair comparison, we evaluate MIDAR and RADA using the same ResNet-50 and ResNet-101 backbones in Table 7. As shown in Table 7, with a stronger ResNet-101 backbone, MIDAR shows an improvement of 0.46 points over RADA with ResNet-101 backbone on the Epic-Kitchens dataset. The favorable performance of MIDAR on the different backbones underscores its effectiveness.

5 Conclusions

In this paper, we tackle the problem of domain-generalized action recognition, which is a challenging, yet relatively under-explored problem. We study a wide spectrum of representational invariance/variance learning which is often beneficial in the context of domain-generalized action recognition. We empirically find that naively learning multiple invariances leads to even inferior domain generalization performance compared to the baseline without learning any representational invariances. To tackle this challenge, we introduce MIDAR, an innovative multi-teacher distillation approach that learns nuanced influence from multiple teachers with distinct representational invariances/variances. We propose an adaptive re-weighting layer to learn such nuanced influence from multiple teachers as well as to incorporate both feature-space and output-space distillation. The empirical results on the challenging EPIC-Kitchens dataset with a moderate size demonstrate that MIDAR generalizes across different domains compared to the existing domain generalization methods. Our future work is overcoming this limitation. We plan to improve MIDAR's adaptability to various data scales. Moreover, we plan to apply MIDAR to Transformer architectures and tailor MIDAR to leverage the representational invariance and variance of Transformers.

Acknowledgement. This work was supported in part by the Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korea Government (MSIT) under grant RS-2024-00353131, RS-2021-II212068 (Artificial Intelligence Innovation Hub), and RS-2022-00155911 (Artificial Intelligence Convergence Innovation Human Resources Development (Kyung Hee University)). Additionally, it was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea Government (MSIT) (No. 2022R1F1A1070997).

References

1. Adriana, R., Nicolas, B., Ebrahimi, K.S., Antoine, C., Carlo, G., Yoshua, B.: Fitnets: hints for thin deep nets. In: ICLR (2015)
2. Arjovsky, M., Bottou, L., Gulrajani, I., Lopez-Paz, D.: Invariant risk minimization. arXiv preprint arXiv:1907.02893 (2019)
3. Arnab, A., Dehghani, M., Heigold, G., Sun, C., Lučić, M., Schmid, C.: Vivit: a video vision transformer. In: ICCV (2021)
4. Bertasius, G., Wang, H., Torresani, L.: Is space-time attention all you need for video understanding? In: ICML (2021)
5. Carlucci, F.M., D'Innocente, A., Bucci, S., Caputo, B., Tommasi, T.: Domain generalization by solving jigsaw puzzles. In: CVPR (2019)
6. Carreira, J., Zisserman, A.: Quo vadis, action recognition? a new model and the kinetics dataset. In: CVPR (2017)
7. Chen, X., He, K.: Exploring simple siamese representation learning. In: CVPR (2021)
8. Cheng, S., Gokhale, T., Yang, Y.: Adversarial bayesian augmentation for single-source domain generalization. In: ICCV (2023)
9. Choi, J., Huang, J.B., Sharma, G.: Self-supervised cross-video temporal learning for unsupervised video domain adaptation. In: ICPR (2022)
10. Choi, J., Sharma, G., Chandraker, M., Huang, J.B.: Unsupervised and semi-supervised domain adaptation for action recognition from drones. In: WACV (2020)
11. Choi, J., Sharma, G., Schulter, S., Huang, J.-B.: Shuffle and attend: video domain adaptation. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J.-M. (eds.) ECCV 2020. LNCS, vol. 12357, pp. 678–695. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-58610-2_40
12. Damen, D., et al.: Scaling egocentric vision: the epic-kitchens dataset. In: ECCV (2018)
13. Dave, I., Gupta, R., Rizve, M.N., Shah, M.: TCLR: temporal contrastive learning for video representation. CVIU **219**, 103406 (2022)
14. Fan, H., et al.: Multiscale vision transformers. In: ICCV (2021)
15. Feichtenhofer, C.: X3d: expanding architectures for efficient video recognition. In: CVPR (2020)
16. Feichtenhofer, C., Fan, H., Malik, J., He, K.: Slowfast networks for video recognition. In: ICCV (2019)
17. Feichtenhofer, C., Fan, H., Xiong, B., Girshick, R., He, K.: A large-scale study on unsupervised spatiotemporal representation learning. In: CVPR (2021)
18. Herzig, R., et al.: Object-region video transformers. In: CVPR (2022)
19. Hinton, G., Vinyals, O., Dean, J.: Distilling the knowledge in a neural network. arXiv preprint arXiv:1503.02531 (2015)
20. Hu, C., et al.: Teacher-student architecture for knowledge distillation: a survey. arXiv preprint arXiv:2308.04268 (2023)
21. Jackson, P.T., Abarghouei, A.A., Bonner, S., Breckon, T.P., Obara, B.: Style augmentation: data augmentation via style randomization. In: CVPR Workshop (2019)
22. Kaplun, G., Malach, E., Nakkiran, P., Shalev-Shwartz, S.: Knowledge distillation: Bad models can be good role models. In: NeurIPS (2022)
23. Khosla, P., et al.: Supervised contrastive learning. In: NeurIPS (2020)
24. Li, H., Pan, S.J., Wang, S., Kot, A.C.: Domain generalization with adversarial feature learning. In: CVPR (2018)
25. Li, Y., et al.: Deep domain generalization via conditional invariant adversarial networks. In: ECCV (2018)
26. Lin, J., Gan, C., Han, S.: Tsm: temporal shift module for efficient video understanding. In: ICCV (2019)

27. Liu, Y., Zhang, W., Wang, J.: Adaptive multi-teacher multi-level knowledge distillation. Neurocomputing **415**, 106–113 (2020)
28. Munro, J., Damen, D.: Multi-modal domain adaptation for fine-grained action recognition. In: CVPR (2020)
29. Nie, X., Li, Y., Luo, L., Zhang, N., Feng, J.: Dynamic kernel distillation for efficient pose estimation in videos. In: ICCV (2019)
30. Park, W., Kim, D., Lu, Y., Cho, M.: Relational knowledge distillation. In: CVPR (2019)
31. Patrick, M., et al.: Keeping your eye on the ball: trajectory attention in video transformers. In: NeurIPS (2021)
32. Qian, R., et al.: Spatiotemporal contrastive video representation learning. In: CVPR (2021)
33. Sahoo, A., Shah, R., Panda, R., Saenko, K., Das, A.: Contrast and mix: temporal contrastive video domain adaptation with background mixing. In: NeurIPS (2021)
34. Simonyan, K., Zisserman, A.: Two-stream convolutional networks for action recognition in videos. In: NeurIPS (2014)
35. Singh, A., et al.: Semi-supervised action recognition with temporal contrastive learning. In: CVPR (2021)
36. Sun, B., Feng, J., Saenko, K.: Correlation alignment for unsupervised domain adaptation. In: Domain Adaptation in Computer Vision Applications, pp. 153–171 (2017)
37. Tong, Y., et al.: Quantitatively measuring and contrastively exploring heterogeneity for domain generalization. In: Proceedings of SIGKDD (2023)
38. Tran, D., Bourdev, L., Fergus, R., Torresani, L., Paluri, M.: Learning spatiotemporal features with 3d convolutional networks. In: ICCV (2015)
39. Volpi, R., Murino, V.: Addressing model vulnerability to distributional shifts over image transformation sets. In: ICCV (2019)
40. Volpi, R., Namkoong, H., Sener, O., Duchi, J.C., Murino, V., Savarese, S.: Generalizing to unseen domains via adversarial data augmentation. In: NeurIPS (2018)
41. Wang, L., et al.: Temporal segment networks for action recognition in videos. TPAMI **41**(11), 2740–2755 (2018)
42. Wang, S., Yu, L., Li, C., Fu, C.-W., Heng, P.-A.: Learning from extrinsic and intrinsic supervisions for domain generalization. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J.-M. (eds.) ECCV 2020. LNCS, vol. 12354, pp. 159–176. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-58545-7_10
43. Wang, Z., Luo, Y., Qiu, R., Huang, Z., Baktashmotlagh, M.: Learning to diversify for single domain generalization. In: ICCV (2021)
44. Wu, C.Y., et al.: Memvit: memory-augmented multiscale vision transformer for efficient long-term video recognition. In: CVPR (2022)
45. Xu, D., Xiao, J., Zhao, Z., Shao, J., Xie, D., Zhuang, Y.: Self-supervised spatiotemporal learning via video clip order prediction. In: CVPR (2019)
46. Yan, S., et al.: Multiview transformers for video recognition. In: CVPR (2022)
47. Yao, Z., Wang, Y., Wang, J., Philip, S.Y., Long, M.: Videodg: generalizing temporal relations in videos to novel domains. TPAMI **44**(11), 7989–8004 (2021)
48. You, S., Xu, C., Xu, C., Tao, D.: Learning from multiple teacher networks. In: Proceedings of SIGKDD (2017)
49. Zhang, H., Chen, D., Wang, C.: Confidence-aware multi-teacher knowledge distillation. In: ICASSP (2022)
50. Zhang, H., Cisse, M., Dauphin, Y.N., Lopez-Paz, D.: mixup: beyond empirical risk minimization. arXiv preprint arXiv:1710.09412 (2017)
51. Zhao, B., Cui, Q., Song, R., Qiu, Y., Liang, J.: Decoupled knowledge distillation. In: CVPR (2022)
52. Zhou, B., Andonian, A., Torralba, A.: Temporal relational reasoning in videos. In: ECCV (2018)

53. Zhou, K., Yang, Y., Qiao, Y., Xiang, T.: Domain generalization with mixstyle. arXiv preprint arXiv:2104.02008 (2021)
54. Zhu, X., Gong, S., et al.: Knowledge distillation by on-the-fly native ensemble. In: NeurIPS (2018)
55. Zou, Y., Choi, J., Wang, Q., Huang, J.B.: Learning representational invariances for data-efficient action recognition. CVIU **227**, 103597 (2023)

ALS-HAR: Harnessing Wearable Ambient Light Sensors to Enhance IMU-Based Human Activity Recognition

Lala Shakti Swarup Ray[1](✉), Daniel Geißler[1], Mengxi Liu[1], Bo Zhou[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. Despite the widespread integration of ambient light sensors (ALS) in smart devices commonly used for screen brightness adaptation, their application in human activity recognition (HAR), primarily through body-worn ALS, is largely unexplored. In this work, we developed ALS-HAR, a robust wearable light-based motion activity classifier. Although ALS-HAR achieves comparable accuracy to other modalities, its natural sensitivity to external disturbances, such as changes in ambient light, weather conditions, or indoor lighting, makes it challenging for daily use. To address such drawbacks, we introduce strategies to enhance environment-invariant IMU-based activity classifications through augmented multi-modal and contrastive classifications by transferring the knowledge extracted from the ALS. Our experiments on a real-world activity dataset for three different scenarios demonstrate that while ALS-HAR's accuracy strongly relies on external lighting conditions, cross-modal information can still improve other HAR systems, such as IMU-based classifiers. Even in scenarios where ALS performs insufficiently, the additional knowledge enables improved accuracy and macro F1 score by up to 4.2 % and 6.4 %, respectively, for IMU-based classifiers and even surpasses multi-modal sensor fusion models in two of our three experiment scenarios. Our research highlights the untapped potential of ALS integration in advancing sensor-based HAR technology, paving the way for practical and efficient wearable ALS-based activity recognition systems with potential applications in healthcare, sports monitoring, and smart indoor environments.

Keywords: Human Activity Recognition · Ambient light sensor · Sensor Fusion · Contrastive Learning · IMU Sensing

1 Introduction

Sensor-based HAR has gained increasing interest in research and industry over the past decade, advocating various sensor modalities like pressure sensors [22,

A. Antonacopoulos et al. (Eds.): ICPR 2024, LNCS 15329, pp. 133–147, 2025.
https://doi.org/10.1007/978-3-031-78110-0_9

33], EMG sensors [13,15], impedance sensors [7,12] and capacitive sensors [8]. Especially with the ubiquity and availability of smart devices, embedded sensors like inertial measurement units (IMUs) [10,14,21] have gained popularity due to their ability to capture motion-related data accurately through great information density.

However, despite the extensive exploration of IMUs in HAR, there is a growing trend towards investigating the potential of other embedded sensors like BLE [4,27], WiFi signals [30,32], temperature sensors [3] and ALS [29], which is presented in this work. Nowadays, ALS is embedded in almost all portable smart devices with a screen primarily used for adaptive screen brightness adjustments based on changing environmental lighting conditions [6]. Such a sensor operates passively without direct user interaction and can be exploited to provide valuable contextual information about the user's surroundings and activities. Additionally, it consumes minimal power, contributing to energy-efficient implementations, particularly on battery-powered devices. For this work, we aim to benefit from the ubiquity of ALS in smart mobile devices, eliminating the need for additional hardware through straightforward accessibility.

To the best of our knowledge, despite the promising advantages and availability, limited research has been done regarding exploring body-worn ALS in HAR as a motion-sensing modality. Throughout related Multi-modal sensor fusion works like [16,26,28], static ALS and other ambient sensors placed in the environment are used for positional understanding for motion localization.

Wearable, body-worn ALS holds promise for HAR, particularly in indoor environments with stable lighting conditions where external factors affecting light intensity are minimal. Xu et al. have exploited the wearable ALS to generate IMU data for improving nursing-based HAR [29]. Similarly, Sadaghiani et al. used wearable photodiodes to gather blood pressure signals of the wearer's body [24]. Despite the promises, these models suffer from the obvious problem of being sensitive to external lighting conditions. When the changes in light conditions are more significant or comparable to those impacted by the user's movement, the model performance drastically decreases, making them useless for such conditions. Even further, just like the overall nature of vision sensors, the ALS can not work in dark environments [29] as stated by Xu et al.

In this paper, we try to solve this problem by investigating different cross-modal approaches that can empower other sensor-based modalities, like IMU-based HAR, by using knowledge transfer techniques from ALS to IMU. Therefore, we aim to maximize the knowledge extracted from ALS even in unfavorable, fluctuating light conditions, improving HAR performance through ALS independently of environmental influences. In summary, the main contribution of this work can be summarized as follows:

- Multi-modal HAR dataset: A novel multi-modal dataset containing nine different activities for a total of 5.28 h performed by 16 participants along three different environmental scenarios gathering right wrist IMU signal, right wrist ALS signal, video footage and SMPL pose synchronized together as visualized in Fig. 1.

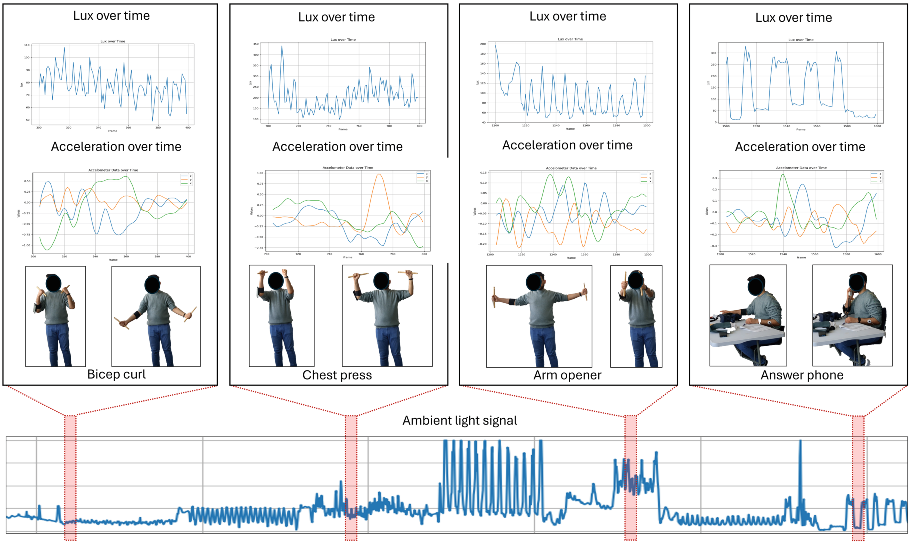


Fig. 1. Synchronized ambient light and IMU accelerometer signals extracted from the collected dataset aligned with the some of the labelled activity classes.

- LightHAR: An activity recognition model based only on the wrist-based ALS with a detailed comparison to wrist-based IMU, 3D pose-based, and video-based activity recognition for the three different scenarios.
- Light embedded InertialHAR: Two different strategies to improve inertial HAR utilizing both ALS and IMU signal during training and only IMU signal for inference.

2 Related Work

2.1 ALS-HAR

HAR is a continuously evolving field that leverages various sensing modalities to identify and monitor human activities. ALS has shown promise in enhancing the accuracy and applicability of HAR systems, especially deployed as external environmental sensors through works like [16] presenting a Deep Convolutional Neural Network to recognize human activities using binary ambient sensors to identify activities of daily living. In [1], the landscape of available sensors for HAR has been analyzed by Ahamed et al., investigating the importance of environmental sensors, especially the ALS, to detect the early signs of dementia in residential care. Integrated into smart wallpapers, multi ALS has been implemented by Shi et al. to recognize human motions with an accuracy of 96% utilizing the information of light reflections gathered through photodiodes hidden inside the wallpaper [26]. Focusing on industrial scenarios and ambient assisted living, Salem et al. have proven the feasibility of fusing the sensor data from

IMU and ALS to achieve an activity recognition performance of 90% across a small set of three classes for each scenario [25].

Environmental Sensing for HAR commonly possesses drawbacks on adaptation to changing light conditions and occlusion of the covered area, wherefore body-worn ALS can be an alternative to enhance HAR performance [9]. Due to their simplicity in operation and low power consumption, they are commonly deployed in consumer wearable devices [20]. In [11], the benefits of low-power ALS have been deployed to harvest energy through photodiodes and simultaneously utilize the ambient light data for self-powered and robust finger gesture detection. Similar work has been done through OptoSense, presenting a novel approach for developing body-worn ALS that is self-powered and capable of being integrated ubiquitously by leveraging photovoltaic cells both to power the sensors and to sense the ambient light, enabling the creation of energy-efficient HAR through light sensors [31]. Similar to the approach of this work, Wang et al. present a multimodal feature fusion model utilizing geomagnetic, ALS, and accelerometer data collected from smartphones to enhance health activity monitoring accuracy in indoor environments by 13.65% compared to classic sensor classification [28].

However, the presented literature barely investigates changing environments and lighting conditions, commonly working in clean and optimal indoor environments, restating the motivation of this work.

2.2 Knowledge Transfer in Sensor HAR

Methods like Don't Freeze [5], Virtual Fusion [17] and Contrastive Left-Right HAR [18] tried using IMU sensors at different positions to improve the overall accuracy of body-worn IMU at specific positions using contrastive learning. Approaches like Multi^{3}Net [23] have tried to improve sensor-based HAR accuracy using other widely available modalities like 3D poses and text embedding. i-Move [12] improved IMU-based HAR using bio-impedance data through contrastive learning.

We believe that the most important use case for ALS data would be empowering other sensor modalities that are more environment invariant through data collected in ideal environments, which we try to achieve through this work.

3 Data Collection

Our experiment encompassed three different scenarios based on different environmental and lighting conditions. It consisted of 16 participants doing 10 activities, including the Null class. The gender distribution was 5 females and 11 males, ages 24 to 35, and weights ranged between 53 kg and 88 kg.

Scenario 1, consisting of subjects 1 to 10, was recorded in a controlled indoor environment with fixed lighting conditions. This environment is ideal for ALS-HAR because of the minimal interference of change in light due to external factors. Scenario 2, consisting of subjects 11 to 13, was recorded in a relatively dark indoor environment with dynamic architectural lights. Most interference in lighting conditions is introduced due to these external factors rather than the motion of the subject itself, making it more challenging than the other two scenarios. Scenario 3, consisting of subjects 14 to 16, was recorded in an outdoor environment during cloudy weather. The clouds and trees moved because wind created small interference in light signals, making it a practical dataset to showcase the usability of ALS-HAR.

Participants are engaged in a series of predetermined activities, including six distinct upper body fitness exercises *boxing, biceps curls, chest press, shoulder, and chest press, arm hold and shoulder press,* and *arm opener* sourced from Pamela's fitness routines available on YouTube ([1]). Additionally, three supplementary hand-focused tasks *sweeping a table, Answering the telephone,* and *wearing a headset* were included, each lasting approximately 20 min.

We used existing consumer-grade devices for data collection to showcase the utility of ALS signals without facing the bottleneck of the new sensor introduction. We utilized a Samsung Galaxy S20 smartphone worn on top of the right wrist with a wristband facing outward, having the same relative position and orientation to the wrist irrespective of the user. This allowed for the collection of both light sensor and IMU data to fulfill the experimental requirements. Data was collected using the Sensor Logger Android application ([2]). Video recordings, captured using a back-facing camera of an iPhone SE, served as supplementary data for annotation purposes.

Sensor Logger automatically synchronizes light and IMU sensor, ensuring a consistent sampling rate of around 30 Hz throughout the session by taking a common time-stamp from the smartphone itself and matching the start and end of the session. The videos collected by a separate smartphone are synchronized manually with the sensor data using a simple trick. At the start and end of each session, the subject needs to do the calibration movement, i.e., fold arms to touch both hands three times to make a unique pattern in the pose and the sensor signal. By mapping these unique patterns of the pose and the sensor signal, we can synchronize both together. Afterward, the videos are manually annotated and can be used directly to annotate the sensor signals.

Typical ALS available in smartphones uses a photodiode, a semiconductor device that generates an electrical current when exposed to light. The intensity of the current is proportional to the amount of light hitting the sensor. The light signal, recorded in lux, is a unit of measurement for illuminance, representing the amount of light per unit area. In this context, lux provided insights into the ambient light conditions surrounding the experiment's environment, especially the light reaching the right wrist based on the subject's movement (Table 1).

[1] https://www.youtube.com/@PamelaRf1.

[2] https://github.com/tszheichoi/awesome-sensor-logger/.

Table 1. Data statistics including subject mass (*kilogram*), height (*centimeter*), gender, and duration of the session (*second*) for the three different scenarios.

Scenario	Subject ID	Age	Height (cm)	Weight (kg)	Gender	Duration (sec)
1: Indoor (Ideal)	1	30	160	53	Male	1394
	2	32	160	53	Female	1466
	3	25	175	65	Female	1399
	4	35	188	88	Male	1362
	5	26	175	86	Male	1376
	6	24	178	85	Female	1487
	7	26	150	50	Female	1385
	8	24	175	80	Male	1482
	9	25	170	65	Male	1393
	10	26	176	55	Male	1319
2: Indoor (Challenging)	11	27	187	85	Male	1225
	12	30	160	53	Male	1240
	13	28	175	65	Male	1127
3: Outdoor	14	26	176	55	Male	1185
	15	35	168	65	Male	1305
	16	33	153	54	Female	1245

4 Method

4.1 LightHAR

We developed a robust ALS-based activity classifier tailored for light sensor data using a 1D bidirectional LSTM-based encoder architecture inspired by the DeepConvLSTM framework [19]. Our model processes input data X with dimensions $(N, 1, 1)$, where N represents the sequence length with unit feature dimension and a single channel. The architecture outputs a probability distribution over 10 classes (9 + null), denoted as $\hat{Y}$.

The model begins with a series of three 1D convolutional layers followed by batch-normalization and dropout layers, each designed for feature extraction. These layers sequentially process the input data to capture relevant patterns and characteristics from the light sensor signals. Each convolutional block is followed by a ReLU activation function, which introduces non-linearity into the model. After feature extraction, the processed data is passed through a bidirectional LSTM layer to capture temporal dependencies in the sequence data. The bidirectional nature of the LSTM layers allows the model to consider both past and future information when making predictions, which enhances the overall performance of the activity classifier. The output from the LSTM layers is then directed through dense layers for classification. The final layer outputs a probability distribution of the 10 classes, enabling the model to determine the most likely activity class for a given sequence of light sensor data. Our architecture focuses on creating a lightweight model with robust and accurate classification based on light sensor inputs (Fig. 2).

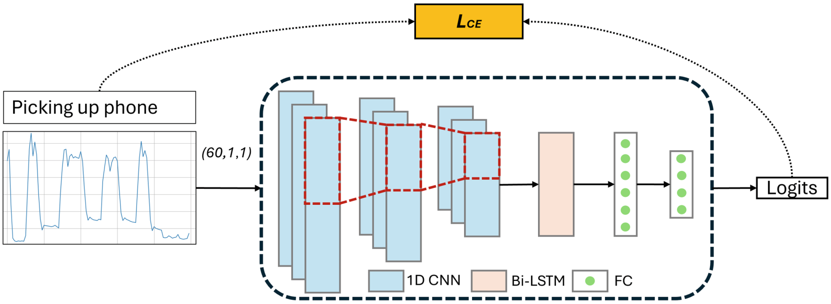


Fig. 2. Overview of the architecture of LightHAR that uses ALS data only for activity classification.

To train the model, we used the cross-entropy loss function, defined as:

$$\mathrm{L_{CE}} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{c=1}^{C} y_{i,c}\log(\hat{y}_{i,c}) \tag{1}$$

where N is the batch size, C is the number of classes, $y_{i,c}$ is the ground truth probability that sample i belongs to class c, and $\hat{y}_{i,c}$ is the predicted probability by the model for class c of sample i.

4.2 Light Embedded InertialHAR

We have designed different strategies for leveraging the knowledge from the ALS modality to enhance the activity recognition accuracy of the IMU modality. As detailed in the Sect. 5.2, ALS, due to its high sensitivity to external light conditions, is susceptible to environmental noise, especially during significant light changes. In contrast, the accelerometer from IMU is known for its environmental robustness and stability. We've developed a variety of strategies that leverage the unique features of both ALS and IMU sensors. These strategies enable us to build a model that only requires the IMU modality during evaluation, effectively mitigating the impact of environmental noise on ALS. This approach is particularly useful in practical scenarios with substantial light fluctuations.

MultiLight InertialHAR. We designed MultiLight InertialHAR by taking inspiration from classic sensor-fusion models that use more than one modality to improve overall HAR accuracy compared to either unimodal system. The model contains two encoders: An ALS encoder partially similar to the LightHAR used for activity classification where the full-connected layers are replaced to generate a dense feature vector of size 256. The IMU encoder also contains a series of 3 1D CNN blocks followed by a bidirectional LSTM and a fully connected layer to generate a dense feature vector of size 256. The extracted features are concatenated afterward and given to a simple classifier consisting of

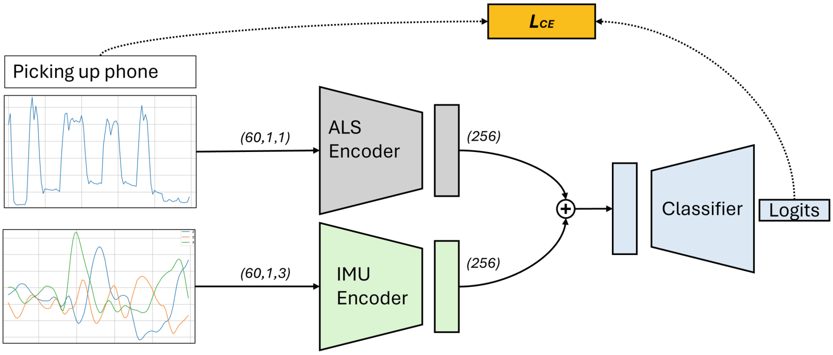


Fig. 3. Overview of MultiLight InertialHAR that takes both ALS and IMU data during training and relies on IMU only during inference by filling the ALS part with zeros as placeholder during inference.

two fully connected layers to map the intermediate features to an activity class as visualized in Fig. 3.

Like the LightHAR model, cross-entropy loss was used to train the model. MultiLight IneritalHAR processes input ALS data $(N, 1, 1)$, and input accelerometer data $(N, 1, 3)$, where N represents the sequence length, 1 is the feature dimension, and 3 is total channels(x, y, z) to output one of the 10 (9+Null) classes.

Since we aim to design a HAR system that utilizes both sensor modalities during the training phase and only the IMU modality during the evaluation phase, we develop a unique data pre-processing pipeline to train the model. Each data point in the dataset is converted to 3 instances: the original and instances where one of the two modalities is replaced by zero, enabling us to evaluate the model even when one is unavailable. During the inference phase, without the presence of ALS data, we can simply $(N, 1, 1)$ input this as a set of 0 and give appropriate data for $(N, 1, 3)$, making it possible to work without changing the architecture.

ContraLight InertialHAR. Inspired by other multi-modal contrastive learning models, We devised another unique strategy where both light and inertial sensor data are used during the training phase, but only inertial sensor data is used during the evaluation phase by utilizing contrastive learning to train this model. The ALS encoder and IMU encoder, identical to those used in MultiLight InertialHAR, extract two feature vectors of size 256. Contrastive loss is then applied to both embeddings based on the original classes to bring representations from the same classes closer together.

The contrastive loss L_{co} is defined as:

$$\mathrm{L}_{\mathrm{co}} = \sum_{i,j} y_{ij} \cdot \max(0, m - \|z_i - z_j\|^2) + (1 - y_{ij}) \cdot \|z_i - z_j\|^2 \tag{2}$$

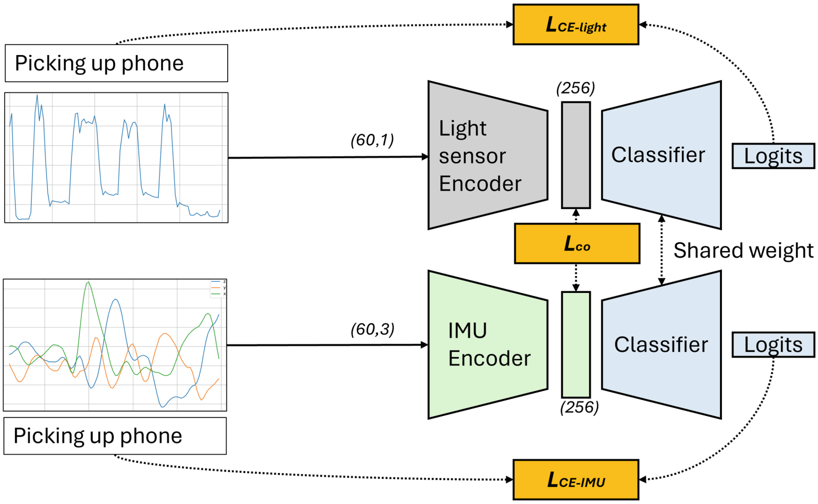


Fig. 4. Overview of ContraLight InertialHAR that takes both ALS and IMU data during training but only IMU during inference.

where z_i and z_j are the feature vectors, y_{ij} is a binary label indicating whether z_i and z_j are from the same class, and m is a margin parameter. Two instances of the same fully connected classifier are utilized with shared weights as visualized in Fig. 4. The overall total loss $\mathcal{L}_{\text{total}}$ is then calculated by summing the contrastive loss and the two cross-entropy losses.

$$\mathrm{L}_{\text{total}} = \mathrm{L}_{\text{co}} + \mathrm{L}_{\text{CE-light}} + \mathrm{L}_{\text{CE-IMU}} \tag{3}$$

We used contrastive loss instead of InfoNCE loss as this is a supervised problem. Therefore, we can directly use the labels as individual clusters instead of the self-supervised clustering task, which is useful in cases where the target activities are different from the source activities.

This approach ensures that during training, the model leverages both $(N, 1, 1)$ ALS and $(N, 1, 3)$ IMU sensor data to learn robust feature representations. However, during inference, we can simply discard the ALS encoder and use the more stable $(N, 1, 3)$ inertial sensor data as input to predict the activity class. Unlike the MultiLight InertialHAR, we do not need to pass a dummy ALS input, making the model even smaller and more simplified without any significant trade-off.

5 Evaluation

5.1 Training Details

To train the models with the collected data, instances were generated using a sliding window technique with a size of 60 (2 sec) and a step of 15 samples (0.5 sec) for both ALS and IMU sensor data. The video and extracted pose, which have 24 frames per second (FPS), are first interpolated to make it 30FPS and afterward sliced accordingly to generate a window of size 60 and a step of 15 frames.

All models are trained using a Nvidia A6000 Ada Lovelace GPU and a Ryzen 5900 processor. Subjects 1 to 7 constituted the training set, while subjects 8 to 10 formed the test set 1 for ideal light conditions. Subjects 11 to 13 formed test set 2 for challenging indoor lighting conditions, and subjects 14 to 16 formed test set 3 for outdoor conditions. Training and validation data were randomly split with a 9:1 ratio during the training process.

The ADAM optimizer, along with a constant learning rate of 0.001, is used to train the model. The models are trained for 300 epochs, and early stopping with a patience of 10 is employed.

5.2 Unimodal Results

To test the effectiveness of ALS as an activity classification modality, we trained other widely used temporal modalities, such as IMU, Pose, and Video, for activity recognition using the same dataset we collected before. All modalities interpolated to have the same sampling frequency and step size. To make them comparable, we made the neural network architecture identical to LightHAR for all other modalities except the input size. Each model was designed with three CNN blocks for feature extraction, a bi-directional LSTM, and two fully connected layers for activity classification, mirroring the structure of LightHAR. As stated in Sect. 4.1, the input of the LightHar is the ALS signal of size $(60, 1, 1)$ while the input for InertialHAR is the IMU signal of size $(60, 1, 3)$. In contrast, the input for the PoseHAR is the extracted SMPL pose from videos using MotionBERT [34] of size $(60, 22, 3)$ with 22 joints, and the input for the VideoHAR is the extracted intermediate video features using Video Vision Transformer (ViViT) [2] of size $(1, 3137, 768)$.

Table 2. Classification accuracy, macro F1, total number of learnable parameters, inference time, and number of Floating Point Operation (FLOP) from an ALS, IMU, and vision-based HAR sharing identical architecture for the three different scenarios.

Modality	Scenario	Accuracy	Macro F1	Parameters	Time (ms)	FLOP
ALS (LightHAR)	1	0.701 ± 0.024	0.639 ± 0.039	**0.294M**	**0.292 ± 0.008**	**25.050M**
	2	0.413 ± 0.017	0.398 ± 0.021			
	3	0.634 ± 0.012	0.608 ± 0.029			
IMU (InertialHAR)	1	0.713 ± 0.041	0.657 ± 0.017	0.360M	0.364 ± 0.023	41.430M
	2	**0.706 ± 0.023**	**0.688 ± 0.023**			
	3	0.722 ± 0.033	0.696 ± 0.017			
Vision (PoseHAR)	1	**0.913 ± 0.029**	**0.896 ± 0.037**	2.425M	0.503 ± 0.017	557.397M
	2	0.658 ± 0.035	0.644 ± 0.017			
	3	**0.0852 ± 0.034**	**0.838 ± 0.016**			
Vision (VideoHAR)	1	0.517 ± 0.025	0.492 ± 0.033	10.322M	0.642 ± 0.029	2531.201M
	2	0.412 ± 0.024	0.396 ± 0.031			
	3	0.366 ± 0.027	0.358 ± 0.022			

As stated in Table 2 we used Accuracy, Macro F1 score, Total number of parameters, Inference time, and Floating Point Operations (FLOP) as metrics to compare all four modalities. For test sets 1 (indoor with fixed external lights) and 3 (outdoor), PoseHAR performed best, followed by IntertialHAR. LightHAR, despite having the fewest learnable parameters and a single channel input, had comparable results to InertialHAR for test set 1. For test set 2 (indoor with dynamic external lights), InertialHAR performed best, followed by PoseHAR and LightHAR. This change can be attributed to the accelerometer being light-invariant and more stable than other vision-based modalities. The VideoHAR, despite having the highest number of learnable parameters, performed worst in all cases, which can be attributed to the feature extracted by ViViT. ViViT extracts video features but has not been specifically trained to extract the features of the person in the video for activity recognition. The extracted features might contain more information about the background rather than the person himself, which is useless for the activity recognition problem, making it useless for this specific use case.

In terms of inference time, LightHAR, with the lowest number of FLOP, has the fastest inference time, followed by InertialHAR, PoseHAR, and VideoHAR. PoseHAR and VideoHAR also take intermediate features as input, so considering the inference time for pose estimation and video feature extraction would make this even higher, making them unsuitable for real-time use cases.

Despite LightHAR's promising inference time and comparable results to InertialHAR for test set 1, it does not solve the underlying problem related to the unreliability of ALS sensors in challenging conditions like test sets 1 and 2. To address this, we developed cross-modal knowledge transfer as described in Sect. 4.2.

5.3 Cross-modal Results

As discussed in the Sect. 5.2, if we consider all scenarios, InterialHAR is more accurate, more reliable/stable, and has a comparable inference time to LightHAR, making it a better choice for activity recognition.

We developed strategies like MultiLight InertialHAR, which takes $(60, 1, 3)$ IMU input and a dummy array of $(60, 1, 1)$ for activity classification. In this strategy, both IMU and light modality from ideal conditions (fixed light indoor) are used for training the model. Similarly, ContraLight InertialHAR takes only $(60, 1, 3)$ IMU input for activity classification, but both IMU and light modality from ideal conditions (fixed light indoor) are used for training the model. The same metrics from before are used to compare all the models.

As we can see in Table 3, for all 3 test sets, both MultiLight InertialHAR and ContraLight InertialHAR outperformed the baseline InertialHAR although requiring the same input $(60, 1, 3)$ accelerometer data during inference phase. The Sensor Fusion model that is identical to MultiLight InertialHAR but requires both $(60, 1, 3)$ IMU input and $(60, 1, 1)$ ALS input outperforms all of them in test set 1 having ideal lighting conditions, but its performance gets even worse than the baseline InterialHAR for test set 2 where the lighting conditions are

Table 3. IMU Classification accuracy, macro F1, total number of learnable parameters, inference time, and number of Floating Point Operations (FLOP) for baseline InertialHAR, Multi-Light InertialHAR, Contra-Light InertialHAR and Sensor Fusion(IMU+ALS).

Model	Scenario	Accuracy	Macro F1	Parameters	Time (ms)	FLOP
InertialHAR (Baseline)	1	0.713 ± 0.041	0.657 ± 0.017	**0.360M**	**0.363 ± 0.023**	**41.430M**
	2	0.706 ± 0.023	0.688 ± 0.023			
	3	0.722 ± 0.033	0.696 ± 0.017			
Multi-Light InertialHAR	1	0.719 ± 0.035	0.681 ± 0.027	0.852M	0.388 ± 0.022	198.216M
	2	0.711 ± 0.023	0.690 ± 0.021			
	3	0.725 ± 0.034	0.696 ± 0.035			
Contra-Light InertialHAR	1	0.755 ± 0.031	0.721 ± 0.038	0.366M	**0.363 ± 0.040**	41.442M
	2	**0.731 ± 0.033**	**0.719 ± 0.018**			
	3	**0.756 ± 0.018**	**0.729 ± 0.029**			
Sensor Fusion (ALS+IMU)	1	**0.858 ± 0.051**	**0.820 ± 0.036**	0.852M	0.391 ± 0.031	198.216M
	2	0.681 ± 0.031	0.669 ± 0.025			
	3	0.723 ± 0.024	0.0693 ± 0.016			

challenging it doesn't provide any improvements for test set 3 either. Regarding inference time, the baseline InertialHAR, having the lowest number of FLOP, performs faster than all other models. The MultiLight InertialHAR requires an additional dummy ALS input during inference, which has much higher number of FLOP and is slower than baseline InertialHAR despite being more accurate. The ContraLight InertialHAR, while not surpassing the baseline, demonstrates a very similar number of FLOP and performs on par with the baseline in terms of speed. This efficiency, combined with its superior accuracy and F1 score compared to both the baseline InertialHAR and MultiLight InertialHAR, makes it a competitive model.

5.4 Discussions

As stated in Sect. 5.2, despite having decent inference time and comparable results compared to other sensor modalities like IMUs, ALS can not be used as a universal Unimodal-HAR. Despite its limited working environments, ALS would make a very good modality for specific use cases in smart indoor environments. For example, hospitals or Care Homes have comparatively stable lighting, and the fast inference, passive sensing, and low-power use of ALS make them suitable for this job.

Also, because of their wide availability in smartphones and smartwatches, they can be used for simple yet repetitive tasks like step counting or other types of fitness activity counters, along with IMU modality through sensor fusion.

As discussed in Sect. 5.3, A large amount of multi-modal activity data can be collected in ideal conditions to enhance other sensor modalities like IMU through our knowledge transfer strategies.

5.5 Limitations

In our current study, we exclusively utilized the Galaxy S20 to collect all data, limiting our insights to a single device's performance. Testing a different device type other than the one used to collect the training data could provide valuable insights into cross-device ALS-HAR reliability, particularly in scenarios where other sensor-based modalities like IMU lag behind. Additionally, employing more than one ALS sensor at different parts of the body could potentially enhance overall accuracy. This approach may provide more robustness against varying light conditions and outdoor environments, a direction we plan to explore in future research to improve the robustness and reliability of our findings.

6 Conclusion

In summary, our study delves into the realm of wearable ALS for HAR, showcasing its potential in understanding human motions. We developed LightHAR, a novel approach utilizing wrist-based ambient light signals for HAR, tested it for different scenarios, and compared it with other commonly used modalities for HAR. By integrating ALS with IMU through sensor fusion and contrastive classification, we enhanced the accuracy of InertialHAR systems. Our light-embedded InertialHAR approach, which relies solely on inertial data during inference, exhibited notable improvements in accuracy compared to traditional IMU-based classifiers. Although our study is promising, it is essential to acknowledge its limitations, and further research is warranted to validate our findings across devices and explore practical applications of ambient light-enhanced HAR systems.

Acknowledgements. The research reported in this paper was supported by the BMBF in the project VidGenSense (01IW21003).

References

1. Ahamed, F., Shahrestani, S., Cheung, H.: Internet of things and machine learning for healthy ageing: identifying the early signs of dementia. Sensors **20**(21), 6031 (2020)
2. Arnab, A., Dehghani, M., Heigold, G., Sun, C., Lučić, M., Schmid, C.: Vivit: a video vision transformer. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 6836–6846 (2021)
3. Demrozi, F., Pravadelli, G., Bihorac, A., Rashidi, P.: Human activity recognition using inertial, physiological and environmental sensors: a comprehensive survey. IEEE Access **8**, 210816–210836 (2020)
4. Demrozi, F., Turetta, C., Chiarani, F., Kindt, P.H., Pravadelli, G.: Estimating indoor occupancy through low-cost BLE devices. IEEE Sens. J. **21**(15), 17053–17063 (2021)

5. Fortes Rey, V., Nshimyimana, D., Lukowicz, P.: Don't freeze: finetune encoders for better self-supervised har. In: Adjunct Proceedings of the 2023 ACM International Joint Conference on Pervasive and Ubiquitous Computing & the 2023 ACM International Symposium on Wearable Computing, UbiComp/ISWC 203 Adjunct, pp. 195–196. Association for Computing Machinery, New York (2023). https://doi.org/10.1145/3594739.3610790
6. Isuwa, S., Amos, D., Singh, A.K., Al-Hashimi, B.M., Merrett, G.V.: Maximising mobile user experience through self-adaptive content-and ambient-aware display brightness scaling. J. Syst. Architect. **145**, 103023 (2023)
7. Jiang, D., Wu, Y., Demosthenous, A.: Hand gesture recognition using three-dimensional electrical impedance tomography. IEEE Trans. Circuits Syst. II Express Briefs **67**(9), 1554–1558 (2020)
8. Kumar, R.P., Melcher, D., Buttolo, P., Jia, Y.: Tracking occupant activities in autonomous vehicles using capacitive sensing. IEEE Trans. Intell. Transport. Syst. **24**, 6800–6819 (2023)
9. Lara, O.D., Labrador, M.: A survey on human activity recognition using wearable sensors. IEEE Commun. Surv. Tutor. **15**, 1192–1209 (2013). https://doi.org/10.1109/SURV.2012.110112.00192
10. Li, H., Derrode, S., Pieczynski, W.: An adaptive and on-line imu-based locomotion activity classification method using a triplet markov model. Neurocomputing **362**, 94–105 (2019)
11. Li, Y., Li, T., Patel, R.A., Yang, X.D., Zhou, X.: Self-powered gesture recognition with ambient light. In: Proceedings of the 31st Annual ACM Symposium on User Interface Software and Technology, pp. 595–608 (2018)
12. Liu, M., Rey, V.F., Zhang, Y., Ray, L.S.S., Zhou, B., Lukowicz, P.: imove: exploring bio-impedance sensing for fitness activity recognition. arXiv preprint arXiv:2402.09445 (2024)
13. Martín-Fuentes, I., Oliva-Lozano, J.M., Muyor, J.M.: Electromyographic activity in deadlift exercise and its variants. a systematic review. PloS One **15**(2), e0229507 (2020)
14. McGrath, J., Neville, J., Stewart, T., Cronin, J.: Upper body activity classification using an inertial measurement unit in court and field-based sports: a systematic review. Proc. Inst. Mech. Engineers Part P: J. Sports Eng. Technol. **235**(2), 83–95 (2021)
15. Mekruksavanich, S., Jantawong, P., Hnoohom, N., Jitpattanakul, A.: Human activity recognition for people with knee abnormality using surface electromyography and knee angle sensors. In: 2023 Joint International Conference on Digital Arts, Media and Technology with ECTI Northern Section Conference on Electrical, Electronics, Computer and Telecommunications Engineering (ECTI DAMT & NCON), pp. 483–487. IEEE (2023)
16. Mohmed, G., Lotfi, A., Pourabdollah, A.: Employing a deep convolutional neural network for human activity recognition based on binary ambient sensor data. In: Proceedings of the 13th ACM International Conference on PErvasive Technologies Related to Assistive Environments, PETRA 2020. Association for Computing Machinery, New York (2020). https://doi.org/10.1145/3389189.3397991
17. Nguyen, D.A., Pham, C., Le-Khac, N.A.: Virtual fusion with contrastive learning for single sensor-based activity recognition. arXiv preprint arXiv:2312.02185 (2023)
18. Nshimyimana, D., Rey, V.F., Lukowic, P.: Contrastive left-right wearable sensors (imus) consistency matching for har. arXiv preprint arXiv:2311.12674 (2023)
19. Ordóñez, F.J., Roggen, D.: Deep convolutional and LSTM recurrent neural networks for multimodal wearable activity recognition. Sensors **16**(1), 115 (2016)

20. Perez, A.J., Zeadally, S.: Recent advances in wearable sensing technologies. Sensors **21**(20), 6828 (2021)
21. Pesenti, M., Invernizzi, G., Mazzella, J., Bocciolone, M., Pedrocchi, A., Gandolla, M.: IMU-based human activity recognition and payload classification for low-back exoskeletons. Sci. Rep. **13**(1), 1184 (2023)
22. Ray, L.S.S., Zhou, B., Suh, S., Lukowicz, P.: Pressim: an end-to-end framework for dynamic ground pressure profile generation from monocular videos using physics-based 3d simulation. In: 2023 IEEE International Conference on Pervasive Computing and Communications Workshops and other Affiliated Events (PerCom Workshops), pp. 484–489. IEEE (2023)
23. Rey, V.F., Ray, L.S.S., Qingxin, X., Wu, K., Lukowicz, P.: Enhancing inertial hand based har through joint representation of language, pose and synthetic imus. arXiv preprint arXiv:2406.01316 (2024)
24. Sadaghiani, S.M., Ardakani, A., Bhadra, S.: Ambient light-driven wireless wearable finger patch for monitoring vital signs from ppg signal. IEEE Sens. J. (2023)
25. Salem, Z., Weiss, A.: Improved spatiotemporal framework for human activity recognition in smart environment. Sensors (Basel, Switzerland) **23**, 132 (2022). https://doi.org/10.3390/s23010132
26. Shi, C., Li, T., Niu, Q.: An intelligent wallpaper based on ambient light for human activity sensing. In: Liu, Z., Wu, F., Das, S.K. (eds.) WASA 2021. LNCS, vol. 12939, pp. 441–449. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-86137-7_47
27. Vesa, A.V., et al.: Human activity recognition using smartphone sensors and beacon-based indoor localization for ambient assisted living systems. In: 2020 IEEE 16th International Conference on Intelligent Computer Communication and Processing (ICCP), pp. 205–212. IEEE (2020)
28. Wang, X., Wang, Y., Wu, J.: Position-aware indoor human activity recognition using multisensors embedded in smartphones. Sensors **24**(11) (2024). https://doi.org/10.3390/s24113367. https://www.mdpi.com/1424-8220/24/11/3367
29. Xu, C., et al.: The visual accelerometer: a high-fidelity optic-to-inertial transformation framework for wearable health computing. In: 2022 IEEE 10th International Conference on Healthcare Informatics (ICHI), pp. 319–329. IEEE (2022)
30. Yadav, S.K., et al.: Csitime: privacy-preserving human activity recognition using wifi channel state information. Neural Netw. **146**, 11–21 (2022)
31. Zhang, D., et al.: Optosense: towards ubiquitous self-powered ambient light sensing surfaces. Proc. ACM Interact. Mobile Wearable Ubiq. Technol. **4**(3), 1–27 (2020)
32. Zhang, J., et al.: Data augmentation and dense-lstm for human activity recognition using wifi signal. IEEE Internet Things J. **8**(6), 4628–4641 (2020)
33. Zhou, B., Suh, S., Rey, V.F., Altamirano, C.A.V., Lukowicz, P.: Quali-mat: evaluating the quality of execution in body-weight exercises with a pressure sensitive sports mat. Proc. ACM Interact. Mob. Wearable Ubiq. Technol. **6**(2), 1–45 (2022)
34. Zhu, W., Ma, X., Liu, Z., Liu, L., Wu, W., Wang, Y.: Motionbert: a unified perspective on learning human motion representations. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 15085–15099 (2023)

Spatio-Temporal Domain-Aware Network for Skeleton-Based Action Representation Learning

Jiannan Hu[1], Cong Wu[1], Tianyang Xu[1], Xiao-Jun Wu[1](✉), and Josef Kittler[2]

[1] School of Artificial Intelligence and Computer Science, Jiangnan University, Wuxi, People's Republic of China
{jiannan_hu,congwu}@stu.jiangnan.edu.cn,
{tianyang.xu,wu_xiaojun}@jiangnan.edu.cn

[2] Centre for Vision, Speech and Signal Processing, University of Surrey, Guildford GU2 7XH, UK
j.kittler@surrey.ac.uk

Abstract. Considering its capability to extract implicit patterns from unlabeled data, contrastive learning has been widely employed in unsupervised skeleton-based action recognition. Spatio-temporal modeling is a key component in understanding skeleton sequence. However, existing methods often adopt rudimentary mechanisms or even completely overlook this aspect, leading to suboptimal performance in downstream tasks. In this paper, we propose a **S**patio-**T**emporal **D**omain-**A**ware **Net**work (STDA-Net). Firstly, the features extracted from the backbone extractor are further decoded into a triple-stream representation, corresponding to the spatial, temporal and global domains, respectively. Following that, an innovative approach named Triple Attention Transformer Module (TATM) is proposed to achieve customized spatio-temporal reasoning. TATM consists of three independent attention modules and a shared feedforward layer, thus achieving reasoning in different domains in a more efficient manner. Finally, domain-aware projectors are used to obtain richer spatio-temporal representations, providing a basis for the subsequent construction of inter-domain and intra-domain contrasts. Comprehensive experiments on NTU-RGB+D 60&120 and PKU-MMD datasets demonstrate the superior performance of STDA-Net.

Keywords: Skeleton-based action recognition · Contrastive Learning · Spatio-temporal Awareness · Triple Attention Transformer

1 Introduction

Human action recognition is an important research field in artificial intelligence [10]. It focuses on the recognition and understanding of the actions performed by one or more individuals through machine learning or deep learning algorithms. Among them, skeleton-based action recognition has gradually emerged with its unique advantages, such as stronger robustness, free from environmental interference [36] compared with RGB-based framework. These advantages have catalyzed a considerable amount of research work [24].

A. Antonacopoulos et al. (Eds.): ICPR 2024, LNCS 15329, pp. 148–163, 2025.
https://doi.org/10.1007/978-3-031-78110-0_10

The exploration of related fields has indeed yielded excellent achievements [2,21,32,34]. However, a problem has emerged: insufficient data. In the past, skeleton action recognition tasks were usually based on supervised paradigms, which required large amounts of labeled data. However, the acquisition and labeling of skeleton data is time-consuming and laborious. Therefore, more attention is being directed towards unsupervised skeleton representation learning [37]. Current unsupervised skeleton-based action recognition mainly includes two paradigms: generative learning and contrastive learning. Contrastive learning, in particular, is a potent framework where the primary concept involves pulling positive pairs closer and pushing negative pairs further apart in a high-dimensional semantic space [5]. This approach enhances the discriminative power of representations learned for skeleton action recognition, which proves advantageous for deciphering intricate action patterns [23]. As a result, contrastive learning has emerged as a leading method in unsupervised skeleton-based action recognition [3].

The earliest work [23] drew on the paradigm from MoCo [11], which introduced the idea of computing similarities between augmented instances of input skeleton sequences to extract inherent patterns from unlabeled data. However, it overlooked spatiotemporal modeling, a crucial step in understanding actions. In contrast, [7] encoded skeleton sequences separately based on temporal and spatial domains. Yet, this complete separation resulted in a coarse understanding of spatiotemporal relationships. In general, most previous methods overlooked the necessity of spatio-temporal domain awareness or relied on very basic strategies. The temporal domain typically provides dynamic information about action sequences, while the spatial domain delineates the position and posture of the human body in three-dimensional space. Modeling their differences and associations contributes to a deeper understanding of human action. Previous approaches' neglect of this aspect significantly restricts the model's ability to effectively comprehend complex patterns in skeleton sequences.

Building on the aforementioned insights, we propose a novel Spatio-Temporal Domain-Aware Network (STDA-Net). While following the paradigm of MoCo, our innovations primarily enhance the encoder and refine the loss calculation. To begin, we incorporate a Graph Convolutional Network (GCN)-based feature extractor at the encoder's outset, drawing from its effective graph perception capabilities [1,29] to derive an initial representation. Subsequently, we develop a triple-stream representation through embedding and fusion modules, tailored to temporal, spatial, and global domains respectively. For efficient cross-domain reasoning, we introduce the Triple Attention Transformer Module (TATM). This module conducts sequence modeling using triple attention layers and employs a unified feedforward layer for information updating. Additionally, complementing intra-domain contrast pairs, we introduce a cross-domain loss to further enhance the model's ability to learn robust, discriminative representations. Finally, domain-aware projectors are utilized to map features into the temporal and spatial domains.

The comparisons with the mainstream methods fully prove the superiority of our structure. The results of the ablation experiment demonstrate the effectiveness of our innovation.

The main contributions of STDA-Net are as follows:

- We form temporal and spatial representations through initial feature extraction. With a subsequent fusion module, we obtain the global representation that incorporates spatiotemporal information, thus forming a triple-stream representation.
- A triple attention transformer module (TATM) is proposed, in which features of different domains are modeled separately to obtain domain-specific representations.
- With domain-aware projectors, we obtain original and domain-aware spatial and temporal outputs. Inter-domain loss is formed between outputs from different domains, and intra-domain loss is formed between outputs from the same domain.
- STDA-Net outperforms the state-of-the-art networks on multiple datasets.

2 Related Work

2.1 Graph Convolutional Network for Skeleton-Based Action Recognition

The human 3D skeleton naturally forms a topological graph, making Graph Convolutional Networks (GCNs) increasingly popular for skeleton-based action recognition. [34] introduced the spatial temporal graph convolutional network (ST-GCN), which models human joints as graph vertices and connectivity and time as edges. This work demonstrated GCN's advantages, sparking further research. [30] expanded local attention to a global scale using relative distance and angle, also introducing a new graph adjacency matrix that improved recognition accuracy. [1] proposed channel-level topology optimized graph convolution (CTR-GC), combining shared and channel-specific topologies for better joint connections. We choose GCN as our primary feature extractor, specifically adopting CTR-GCN due to its strong capabilities.

2.2 Contrastive Networks for Unsupervised Skeleton-Based Action Recognition

Contrastive learning is widely used in skeleton action recognition due to its ability to learn implicit patterns from unlabeled data. In contrastive learning of skeleton data, each skeleton instance is typically augmented into two augmented instances, and the encoder is then trained to generate discriminative features, making instances of the same skeleton more similar in representation than instances of different skeletons. A large amount of work has been done to improve the contrast process. In [9], extreme augmentations were introduced to acquire diverse positive samples, and new motion patterns are brought to improve the generality of the learned representation. In [7], features of different granularities were acquired through the down-sampling operation, enabling the contrasts of features at multiple-level. Based on the spatial structure of the human skeleton, [18] partitioned it into static and dynamic regions. They applied different data transformations to each region to achieve adaptive modeling. The above work have demonstrated the effectiveness of contrastive learning in the task of skeleton action recognition.

2.3 Spatio-Temporal Modeling Network

Skeleton data inherently reflects the spatiotemporal characteristics of human motion. Therefore, spatial and temporal modeling are crucial in skeleton action recognition. In [31], spatial-specific features and temporal-specific features were extracted and modeled respectively, and used for loss calculation. Unfortunately, it only adopted a simple decoupling mechanism to obtain spatiotemporal features, resulting in representations that lacked discriminative power. [7] modeled temporal and spatial features separately, and considered the contrasts between features of different granularities to obtain more discriminative representations. But previous work [13,17] has demonstrated that completely independent modeling of spatial and temporal is suboptimal for skeleton action recognition.

3 Methodology

3.1 The Overall Framework

Our model paradigm is inspired by MoCo [11], a robust approach for unsupervised representation learning. MoCo initially encodes input data into feature vectors, which are then contrasted with feature vectors from a dynamically updated queue of negative samples. This contrastive process aims to maximize the similarity between positive sample pairs while minimizing similarity with negative samples, thereby enhancing model robustness.

The overall framework of our model is illustrated in Fig. 1(a). Given an input skeleton sequence x, the process begin with standard data augmentation techniques such as rotation, cropping, and flipping. Subsequently, we implement block-based masking in both temporal and spatial dimensions. This strategy involves grouping consecutive frames and skeleton joints into blocks, where temporal masking involves simultaneously masking features from consecutive frames, and spatial masking involves masking connected joints. This block-based masking strategy effectively prevents information leakage in the skeleton sequence processing [28], resulting in the generation of query sample x_q and the key sample x_k. These samples x_q and x_k are then processed by respective encoders and projectors. The encoders transform the skeleton data into hidden representations, while the projectors map these representations into an output space suitable for contrastive learning. In downstream tasks, the pre-trained encoder q extracts feature representations from input data, which are subsequently fed into a classifier for action recognition, as depicted in Fig. 1(b).

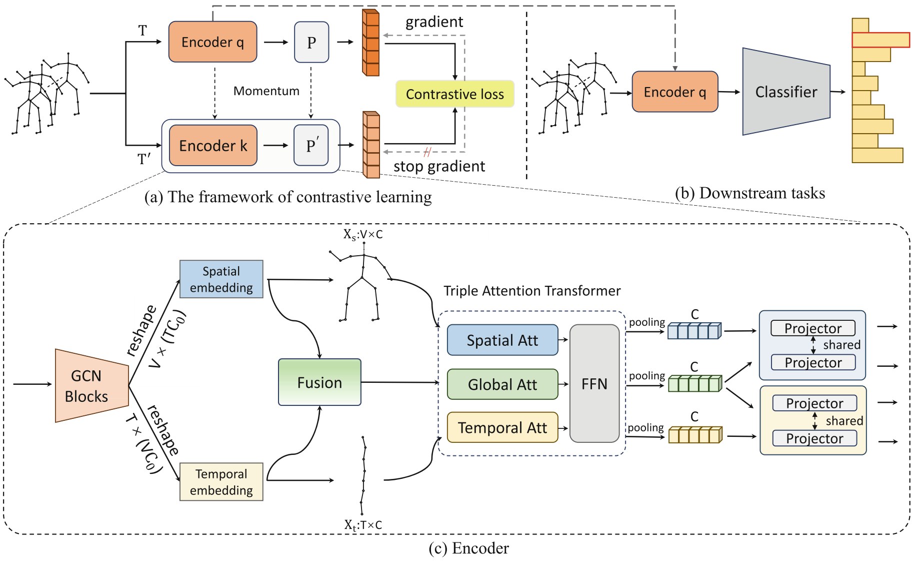


Fig. 1. The framework of contrastive learning (a), illustration for downstream tasks (b) and our proposed encoder structure (c). We show only one encoder in the framework of contrastive learning, another momentum encoder has the same structure as it.

In the following sections, we will provide a detailed description of the proposed encoder's structure, as well as the projection and contrast processes.

3.2 The Proposed Triple-Stream Encoder

Taking encoder q as an example, the specific structure of the encoder is shown in Fig. 1(c). The dimension of the augmented sequence is $x_q \in \mathbb{R}^{3\times T\times V}$.

We use the L-layer graph convolution structure [1] to process the augmented sequence x_q. Through this, x_q is mapped from the coordinate space to the feature space and the feature dimension is $\hat{x} \in \mathbb{R}^{C_l\times T\times V}$, in which C_l, T, V represent the number of channels, frames, and joints. Next, different reshaping operations merge the temporal dimension and spatial dimension with the channel dimension respectively to obtain $\hat{x_t} \in \mathbb{R}^{T\times(VC_l)}$ and $\hat{x_s} \in \mathbb{R}^{V\times(TC_l)}$. The following embedding operations are used to obtain representations corresponding to the temporal and spatial domains respectively. The specific formulas are as follows:

$$\begin{aligned} x_t &= W_1(\sigma(\mathrm{LN}(W_t\hat{x_t} + b_t))) + b_1 \\ x_s &= W_2(\sigma(\mathrm{LN}(W_s\hat{x_s} + b_s))) + b_2 \end{aligned} \tag{1}$$

where $W_t \in \mathbb{R}^{C\times(VC_l)}$, $W_s \in \mathbb{R}^{C\times(TC_l)}$, $W_1 \in \mathbb{R}^{C\times C}$ and $W_2 \in \mathbb{R}^{C\times C}$ are the weight matrix for feature mapping, $b_t \in \mathbb{R}^C$, $b_s \in \mathbb{R}^C$, $b_1 \in \mathbb{R}^C$ and $b_2 \in \mathbb{R}^C$ represent bias, σ represents the activation function ReLU, and LN is short for layer normalization. Therefore, we obtain the temporal and spatial features x_t, x_s respectively.

Next we fuse them to produce a global representation that contains spatiotemporal information, so as to conduct subsequent modeling from a global perspective. The specific operations of the fusion module are summarized as follows:

$$x_g = W_3(\text{Maxpooling}(x_t) || \text{Maxpooling}(x_s)) + b_3 \tag{2}$$

Here, the max pooling is performed on the T dimension and the V dimension respectively. $||$ denotes the concatenation operation. $W_3 \in \mathbb{R}^{C \times C}$ is the weight matrix for feature mapping, and $b_3 \in \mathbb{R}^C$ represents bias.

We argue that, after the GCN encoder performs preliminary processing on the skeleton sequence, the intensity of key features has been initially enhanced. To consolidate these enhancements, we employ maximum pooling to extract and fuse the most prominent spatio-temporal features into a unified global representation. This process yields the features x_s, x_g, and x_t across the three domains.

Following that, these features are passed through a transformer encoder to extract high-dimensional features. In this process, we believe that features from different domains should be modeled independently for reasoning. Therefore, we use Triple Attention Transformer Module (TATM) to process different features. Different attention modules are used to model three features respectively, and then update the information in a shared feedforward layer. Finally, through layer normalization and other operations, we obtain the processed features y_s, y_g, y_t:

$$\begin{aligned} x_s^{'} &= \text{SpatialAttention}(x_s) \\ x_g^{'} &= \text{GlobalAttention}(x_g) \\ x_t^{'} &= \text{TemporalAttention}(x_t) \\ y_s, y_g, y_t &= \text{Feedforward}(x_s^{'}, x_g^{'}, x_t^{'}) \end{aligned} \tag{3}$$

Through the above design, we achieve independent reasoning and modeling in the temporal and spatial domains. Simultaneously, there is a certain degree of exchange of spatiotemporal information when reasoning about global features containing such information. Finally, all features are integrated into a more comprehensive representation through a shared feedforward layer.

3.3 Inter-domain Loss and Intra-domain Loss

Through the aforementioned feature extraction process, we obtained representations in the temporal, spatial, and global domains. Next, we utilize temporal-aware projectors and spatial-aware projectors to map them into the temporal and spatial domains, respectively. Each projector consists of two fully connected layers with an intermediate activation function. The specific mapping process is

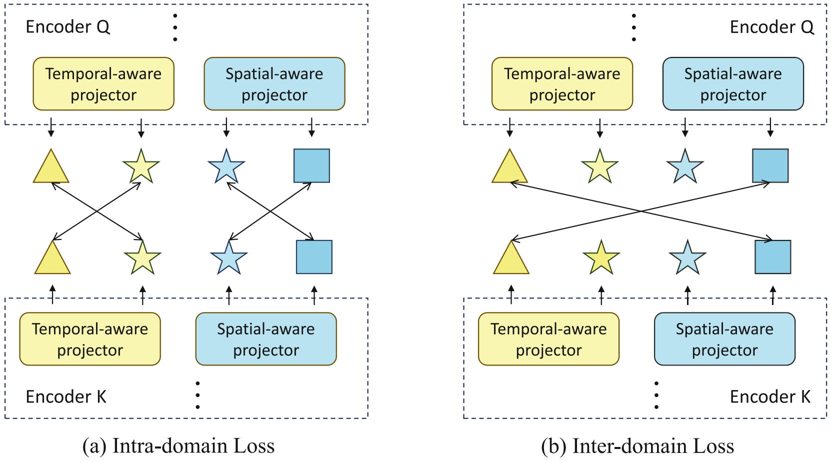


Fig. 2. Intra-domain loss (a) and Inter-domain loss (b). Different symbols represent the outputs generated by the features of different domains through the projectors.

detailed as follows:

$$\begin{aligned} q_s &= \text{S-Projector}(y_s) \\ q_t &= \text{T-Projector}(y_t) \\ q_s^a &= \text{S-Projector}(y_g) \\ q_t^a &= \text{T-Projector}(y_g) \end{aligned} \tag{4}$$

The symbol labeled with superscript “a” denotes the output generated through projection of the global representation. Similarly, the outputs k_s, k_t, k_s^a, k_t^a of encoder k can also been generated follow a same process.

Next, the losses are calculated. Here we introduce InfoNCE loss [22] to measure the distance between outputs. The calculation process is as follows:

$$\mathcal{L}(q,k) = -\log \frac{\mathrm{F}(q,k)}{\mathrm{F}(q,k) + \sum_{m \in M} \mathrm{F}(q,m)} \tag{5}$$

$\mathrm{F}(q,k) = \exp(q \cdot k/\tau)$, τ is the temperature hyperparameter, M represent the queue where the negative samples are.

Intra-domain Loss. The setup of our contrast pairs is shown in Fig. 2. First, we calculate the loss in the spatial domain or the temporal domain. For simplicity, we do not show the calculation process of InfoNCE. The intra-domain loss is as follows:

$$\mathcal{L}_{intra} = \mathcal{L}(q_s, k_s^a) + \mathcal{L}(q_s^a, k_s) + \mathcal{L}(q_t, k_t^a) + \mathcal{L}(q_t^a, k_t) \tag{6}$$

We use the mapped domain-aware features to perform contrast with the original domain features to measure the similarity of samples within the temporal and spatial domains respectively.

Inter-domain Loss. In addition to intra-domain loss calculation, we have also introduced inter-domain loss calculation. The calculation process is as follows

$$\mathcal{L}_{inter} = \mathcal{L}(q_s, k_t) + \mathcal{L}(q_t, k_s) \tag{7}$$

Through the calculation of inter-domain loss, the contrast pairs across temporal and spatial are established, so that the discriminative ability of the contrastive learning framework is significantly enhanced compared with the contrast method in a single domain.

Finally, we set trainable weights for the two losses, and the total loss is expressed as follows:

$$\mathcal{L}_{total} = \lambda_1 \cdot \mathcal{L}_{intra} + \lambda_2 \cdot \mathcal{L}_{inter} \tag{8}$$

4 Experiments

4.1 Dataset

NTU-RGB+D 60. NTU-RGB+D 60 (NTU-60) [25] contains a total of 56,880 samples of 60 types of actions. The skeleton data contains the 3D positions of 25 major body joints per frame. It has two evaluation metrics: Cross-Subject (X-Sub) and Cross-View (X-View). Cross-Subject divides the training set and test set based on person IDs. The training set contains 40,320 samples and the test set contains 16,560 samples. Cross-View divides the training and test sets by camera. The samples captured by camera 1 are used as the test set, and by cameras 2 and 3 are used as the training set. The number of samples is 18,960 and 37,920 respectively.

NTU-RGB+D 120. NTU-RGB+D 120 (NTU-120) [20] is a expansion to NTU-60, which covers all previous data and 60 categories are introduced for a total sample size of 114,480. In terms of evaluation, except the Cross-Subject used in NTU-60, NTU-120 introduces the cross-setup (X-Set) evaluation metric, which uses data collected by even-numbered cameras as the training set and data collected by odd-numbered cameras as the test set.

PKU-MMD. PKU-MMD [19] consists of 1,076 unedited video sequences of 66 participants taken from three views. PKU-MMD includes 51 action categories: 41 daily actions (drinking, waving hand, putting on the glass, etc.) and 10 interaction actions (hugging, shaking hands, etc.). Similar to the NTU RGB+D dataset, this dataset also has two recommended evaluation metrics, Cross-Subject and Cross-View. Following previous methods [7], only the Cross-Subject evaluation metric is adopted in the following experiments.

4.2 The Experimental Setup

In the unsupervised training process, the unlabeled training set is used. We use the stochastic gradient descent algorithm (SGD) as the optimizer. The network is trained on a single NVIDIA 3090 GPU with a batch size of 64 and a learning rate of 0.01 for a total of 450 epochs. The Nesterov momentum is set to 0.9, and the weight decay is set to 1e–4 for NTU RGB+D 60, NTU RGB+D 120, and PKU-MMD I, and 1e–3 for PKU-MMD II.

For downstream tasks, we use labeled data to measure the effect of our upstream unsupervised clustering. We freeze the parameters of the encoder, train the classifier with a learning rate of 0.1 for 80 epochs, and reduce the learning rate to one-tenth of the original at the 50th and 70th epochs, respectively. Specifically, in the transfer learning task, we fine-tune the entire network (including the encoder and classifier) with the initial learning rate set to 0.01.

Table 1. Comparison with mainstream methods on action classification.

Method	NTU-60		NTU-120		PKU-MMD I	PKU-MMD II
	x-sub	x-view	x-sub	x-set	x-sub	x-sub
LongT GAN [37](AAAI'18)	52.1	56.4	–	–	67.7	26.5
PCRP [33](TMM'21)	54.9	63.4	43.0	44.6	–	–
EnGAN-PoseRNN [15](WACV'19)	68.6	77.8	–	–	–	–
H-Transformer [4](ICME'21)	69.3	72.8	–	–	–	–
CrosSCLR [16](CVPR'21)	72.9	79.9	–	–	84.9*	21.2*
AimCLR [9](AAAI'22)	74.3	79.7	63.4	63.4	87.8*	38.5*
Colorization [35](ICCV'21)	75.2	83.1	–	–	–	–
GL-Transformer [14](ECCV'22)	76.3	83.8	66.0	68.7	–	–
ISC [27](ACM MM'21)	76.3	85.2	67.1	67.9	80.9	36.0
HYSP [8](ICLR'23)	78.2	82.6	61.8	64.6	83.8	–
ActCLR [18](CVPR'23)	80.9	86.7	69.0	70.5	–	–
HiCo-Transformer [7](AAAI'23)	81.1	88.6	72.8	74.1	89.3	49.4
STDA-Net	**85.7**	**90.5**	**76.8**	**79.4**	**92.5**	**54.6**

4.3 Comparison to the State-of-the-Art

We compare the proposed STDA-Net with previous mainstream unsupervised methods on several downstream tasks, including action recognition, action retrieval, and transfer learning.

Action Recognition. Action classification task is the most representative downstream task. To verify the validity of STDA-Net, we compared it with previous methods on four benchmarks, including NTU 60&120 and PKU-MMD I&II.

The experimental results are summarised in Table 1. On the NTU-60, STDA-Net achieves a outstanding performance on both metrics, with 85.7% and 90.5%. For NTU-120, our model outperforms previous state-of-the-art methods, achieving excellent results of 76.8% and 79.4% on X-Sub and X-Set, respectively. The PKU-MMD dataset is more challenging due to noise caused by various reasons, but STDA-Net still achieves excellent performance, improving by 3.2% and 5.2% over SOTA on PKU-MMD I and II respectively.

Action Retrieval. In this task, we use KNeighborsClassifier [6] as the classifier. Table 2 shows the results on the NTU dataset. In these benchmarks, our model achieved improvements on most metrics. On the X-Sub of the NTU-60, the accuracy increased by 3.9%. On the X-Sub and X-Set metrics of NTU-120, the increases are 3.2% and 4.8% respectively.

Table 2. Comparison with mainstream methods on action retrieval.

Method	NTU-60		NTU-120	
	x-sub	x-view	x-sub	x-view
LongT GAN [37]	39.1	48.1	31.5	35.5
P&C [26]	50.7	76.3	39.5	41.8
AimCLR [9]	62.0	–	–	–
ISC [27]	62.5	82.6	50.6	52.3
HiCo-Transformer [7]	68.3	84.8	56.6	59.1
SkeAttnCLR [12]	69.4	76.8	46.7	58.0
STDA-Net	**73.3**	**82.5**	**59.8**	**63.9**

Table 3. Comparison with mainstream methods on transfer learning.

Method	Transfer to PKU-MMD II	
	PKU-MMD I	NTU-60
LongT GAN [37]	43.6	44.8
ISC [27]	45.1	45.9
HiCo-Transformer [7]	53.4	56.3
STDA-Net	**61.0**	**66.5**

Transfer Learning. Transfer learning improves learning efficiency and performance by applying already learned knowledge to new problems or scenarios. In this work, we migrate pre-trained models on NTU RGB+D 60 and PKU MMD I to the PKU MMD II dataset. The results presented in Table 3 demonstrate that our STDA-Net brings a performance improvement of 7.6% and 10.2% compared with the current SOTA results.

Experiments on Different Modalities. In all of our experiments, joint data are used as the default input. Here we also employ STDA-Net on different modalities, such as bone and motion. The results are shown in Table 4. When the bone representation is used as the input, the accuracy of our model is 85.1%, and when the input is the motion representation, the accuracy is 82.8%.

Table 4. Compared with the mainstream methods on different modalities.

Method	Joint	Bone	Motion
CrosSCLR [16]	72.9	75.2	72.7
AimCLR [9]	74.3	73.2	66.8
HiCo-GRU [7]	80.6	80.3	78.2
HiCo-LSTM [7]	81.4	81.0	78.9
HiCo-Transformer [7]	81.1	80.3	76.2
STDA-Net	**85.7**	**85.1**	**82.8**

Table 5. The effectiveness of proposed key innovations.

Network	Param(M)	FLOPs(G)	Accuracy
ST Network	49.8	4.17	81.9
STDA-Net w/ shared transformer	37.1	4.20	83.4
STDA-Net w/ separate transformer	62.4	4.20	84.4
STDA-Net w/o intra-loss	–	–	83.0
STDA-Net w/o inter-loss	–	–	84.4
STDA-Net	58.2	4.31	**85.7**

4.4 Ablation Study

Effectiveness of Proposed Innovations. We compare the proposed method with the following variants:

(1) **ST Network**: A two-stream network comprising temporal and spatial streams, with the final contrast constrained to either the spatial or temporal domain.
(2) **STDA-Net with shared transformer**: In this variant, the TATM is substituted with a single transformer module that includes an attention mechanism and a feedforward layer.
(3) **STDA-Net with separate transformer**: Here, the TATM in STDA-Net is replaced by three individual transformer modules, each module consisting of an attention mechanism and a feedforward layer.
(4) **STDA-Net without intra-loss**: This variant of STDA-Net removes the intra-domain loss component, focusing only on inter-domain loss.
(5) **STDA-Net without inter-loss**: Conversely, this variant removes the inter-domain loss component, focusing only on intra-domain loss.

First, we use the two-stream spatiotemporal model (ST Network) for comparative experiments. It represents a simple baseline for modeling completely separate spatial and temporal. By comparison, our three-stream representation network has better results with little increase in model complexity. This also shows that while modeling spatiotemporal differentiation, it is equally important to evaluate actions from the perspective of the entire sequence.

Next, we conduct ablation experiments on the proposed TATM. We replace TATM with a shared transformer module or three independent transformer modules. The experiments show that although the shared transformer module has an advantage in the number of parameters, it leads to a decrease in performance. Specifically, this is because the shared transformer module confuses the independent feature reasoning between different domains. In addition, completely independent transformer modules do not bring any performance improvement, which also shows that there is actually a potential correlation between the features of the three domains. The shared feedforward layer in our TATM integrates the information of different domains to obtain a more comprehensive representation.

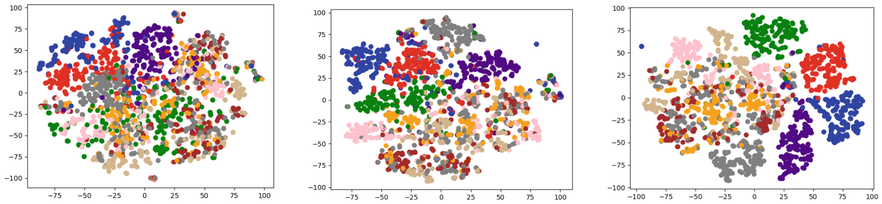
Finally, we conduct experiments by removing the intra-domain and inter-domain losses separately from STDA-Net. This aims to demonstrate their combined effectiveness when used together in the model. Indeed, the experimental results also demonstrate that the application of inter-domian loss and intra-domain loss in our model is natural yet effective.

Table 5 evaluates and compares each experimental configuration. These analyses provide insight into how different components and configurations affect the performance of the model.

Table 6. The exploration of the dimensions of attention computation is performed for each attention module in the proposed TATM.

Attention dimension			Accuracy
Spatial	Global	temporal	
1024	1024	1024	84.3
2048	1024	1024	85.2
1024	2048	1024	84.8
1024	1024	2048	84.7
1536	1024	1024	**85.7**
2560	1024	1024	84.7
3072	1024	1024	85.3

Exploration on the Dimensions of TATM. We explore the influence of different dimensions of attention on the experimental results, as shown in Table 6. First, we increase each attention dimension from 1024 to 2048 respectively to



(a) Original data; DBI=6.7 (b) ST Network; DBI=4.4 (c) STDA-Net; DBI=3.5

Fig. 3. t-SNE visualisation. We randomly select ten categories for visualization. The dots of the same color represent samples belonging to the same action.

explore which dimension has a deeper impact on the final feature expression. Experiments prove that changes in spatial dimensions are more critical, so we conduct experiments on spatial dimensions. With the other two dimensions unchanged, we achieve the best results when the number of channels in the spatial dimension is 1536. But in reality, we fixed the sizes of the temporal and global dimensions, and increasing them moderately may lead to better results.

Table 7. The exploration of encoder parameters of the model.

GCN		Transformer		Accuracy
Layers	Channels	Layers	Heads	
2	64	1	16	85.0
3	**64**	**1**	**16**	**85.7**
4	64	1	16	85.4
3	32	1	16	85.4
3	128	1	16	85.2
3	64	2	16	85.6
3	64	1	4	84.8
3	64	1	8	85.2

Exploration of Encoder Parameters. In addition to the above explorations, we also explore the other parameter settings of the model, and the results are shown in Table 7. The best results are achieved when the GCN encoder has 3 layers and 64 channels, and the transformer encoder has 1 layers and 16 heads. There is no significant change in performance as the model parameters change, indicating the stability of our method.

Visualization. As shown in Fig. 3, we visualize the extracted feature from the unsupervised pre-training. From left to right are the t-SNE visualizations of the raw data, the ST Network, and our STDA-Net. In addition, we quantify the clustering effect using the DBI index, which stands for the Davies-Bouldin Index, offerring a quantitative assessment of clustering effectiveness. The visualization and DBI index prove the clustering effect of our model.

5 Conclusion

In this paper, we present an innovative triple-stream unsupervised contrastive learning framework, STDA-Net. By performing feature extraction and fusion, we obtain features corresponding to spatial, temporal, and global domains. Furthermore, by separately modeling features within each domain, TATM captures domain-specific features. Subsequently, a shared feedforward layer is used to update these features. Finally, inter-domain and intra-domain losses are employed to construct more diverse and challenging contrasts, thereby enhancing the model's discriminative capability. Experimental results across various downstream tasks demonstrate the superiority and robustness of our model.

References

1. Chen, Y., Zhang, Z., Yuan, C., Li, B., Deng, Y., Hu, W.: Channel-wise topology refinement graph convolution for skeleton-based action recognition. In: ICCV, pp. 13359–13368 (2021)
2. Cheng, K., Zhang, Y., He, X., Chen, W., Cheng, J., Lu, H.: Skeleton-based action recognition with shift graph convolutional network. In: CVPR, pp. 183–192 (2020)
3. Cheng, Y.B., Chen, X., Chen, J., Wei, P., Zhang, D., Lin, L.: Hierarchical transformer: unsupervised representation learning for skeleton-based human action recognition. In: 2021 IEEE ICME, pp. 1–6. IEEE (2021)
4. Cheng, Y.B., Chen, X., Chen, J., Wei, P., Zhang, D., Lin, L.: Hierarchical transformer: unsupervised representation learning for skeleton-based human action recognition. In: 2021 IEEE International Conference on Multimedia and Expo (ICME), pp. 1–6 (2021). https://doi.org/10.1109/ICME51207.2021.9428459
5. Chuang, C.Y., Robinson, J., Lin, Y.C., Torralba, A., Jegelka, S.: Debiased contrastive learning. In: NeurIPS, vol. 33, pp. 8765–8775 (2020)
6. Cover, T., Hart, P.: Nearest neighbor pattern classification. IEEE Trans. Inf. Theory **13**(1), 21–27 (1967)
7. Dong, J., Sun, S., Liu, Z., Chen, S., Liu, B., Wang, X.: Hierarchical contrast for unsupervised skeleton-based action representation learning. In: AAAI, vol. 37, pp. 525–533 (2023)
8. Franco, L., Mandica, P., Munjal, B., Galasso, F.: Hyperbolic self-paced learning for self-supervised skeleton-based action representations. arXiv (2023)
9. Guo, T., Liu, H., Chen, Z., Liu, M., Wang, T., Ding, R.: Contrastive learning from extremely augmented skeleton sequences for self-supervised action recognition. In: AAAI, vol. 36, pp. 762–770 (2022)
10. Gupta, N., Gupta, S.K., Pathak, R.K., Jain, V., Rashidi, P., Suri, J.S.: Human activity recognition in artificial intelligence framework: a narrative review. Artif. Intell. Rev. **55**(6), 4755–4808 (2022)

11. He, K., Fan, H., Wu, Y., Xie, S., Girshick, R.: Momentum contrast for unsupervised visual representation learning. In: CVPR, pp. 9729–9738 (2020)
12. Hua, Y., et al.: Part aware contrastive learning for self-supervised action recognition. arXiv (2023)
13. Kay, W., et al.: The kinetics human action video dataset. arXiv (2017)
14. Kim, B., Chang, H.J., Kim, J., Choi, J.Y.: Global-local motion transformer for unsupervised skeleton-based action learning. In: ECCV, pp. 209–225. Springer, Heidelberg (2022). https://doi.org/10.1007/978-3-031-19772-7_13
15. Kundu, J.N., Gor, M., Uppala, P.K., Radhakrishnan, V.B.: Unsupervised feature learning of human actions as trajectories in pose embedding manifold. In: 2019 IEEE WACV, pp. 1459–1467. IEEE (2019)
16. Li, L., Wang, M., Ni, B., Wang, H., Yang, J., Zhang, W.: 3d human action representation learning via cross-view consistency pursuit. In: CVPR, pp. 4741–4750 (2021)
17. Lin, J., Gan, C., Han, S.: Tsm: temporal shift module for efficient video understanding. In: ICCV, pp. 7083–7093 (2019)
18. Lin, L., Zhang, J., Liu, J.: Actionlet-dependent contrastive learning for unsupervised skeleton-based action recognition. In: CVPR, pp. 2363–2372 (2023)
19. Liu, C., Hu, Y., Li, Y., Song, S., Liu, J.: Pku-mmd: a large scale benchmark for continuous multi-modal human action understanding. arXiv (2017)
20. Liu, J., Shahroudy, A., Perez, M., Wang, G., Duan, L.Y., Kot, A.C.: Ntu rgb+ d 120: a large-scale benchmark for 3d human activity understanding. IEEE TPAMI **42**(10), 2684–2701 (2019)
21. Obinata, Y., Yamamoto, T.: Temporal extension module for skeleton-based action recognition. In: 2020 25th International Conference on Pattern Recognition (ICPR), pp. 534–540 (2021). https://doi.org/10.1109/ICPR48806.2021.9412113
22. Oord, A.V.D., Li, Y., Vinyals, O.: Representation learning with contrastive predictive coding. arXiv (2018)
23. Rao, H., Xu, S., Hu, X., Cheng, J., Hu, B.: Augmented skeleton based contrastive action learning with momentum lstm for unsupervised action recognition. Inf. Sci. **569**, 90–109 (2021)
24. Ren, B., Liu, M., Ding, R., Liu, H.: A survey on 3d skeleton-based action recognition using learning method. Cyborg Bionic Syst. **5**, 0100 (2020)
25. Shahroudy, A., Liu, J., Ng, T.T., Wang, G.: Ntu rgb+ d: a large scale dataset for 3d human activity analysis. In: CVPR, pp. 1010–1019 (2016)
26. Su, K., Liu, X., Shlizerman, E.: Predict & cluster: unsupervised skeleton based action recognition. In: CVPR, pp. 9631–9640 (2020)
27. Thoker, F.M., Doughty, H., Snoek, C.G.: Skeleton-contrastive 3d action representation learning. In: ACM MM, pp. 1655–1663 (2021)
28. Tong, Z., Song, Y., Wang, J., Wang, L.: Videomae: masked autoencoders are data-efficient learners for self-supervised video pre-training. In: NIPS, vol. 35, pp. 10078–10093 (2022)
29. Wu, C., Wu, X.J., Kittler, J.: Graph2net: perceptually-enriched graph learning for skeleton-based action recognition. IEEE Trans. Circuits Syst. Video Technol. **32**(4), 2120–2132 (2022). https://doi.org/10.1109/TCSVT.2021.3085959
30. Xing, H., Burschka, D.: Skeletal human action recognition using hybrid attention based graph convolutional network. In: 2022 26th International Conference on Pattern Recognition (ICPR), pp. 3333–3340. IEEE (2022)
31. Xu, B., Shu, X., Zhang, J., Dai, G., Song, Y.: Spatiotemporal decouple-and-squeeze contrastive learning for semisupervised skeleton-based action recognition. IEEE TNNLS (2023)

32. Xu, J., Tasaka, K., Yanagihara, H.: Beyond two-stream: skeleton-based three-stream networks for action recognition in videos. In: 2018 24th International Conference on Pattern Recognition (ICPR), pp. 1567–1573. IEEE (2018)
33. Xu, S., Rao, H., Hu, X., Cheng, J., Hu, B.: Prototypical contrast and reverse prediction: unsupervised skeleton based action recognition. IEEE TMM **25**, 624–634 (2021)
34. Yan, S., Xiong, Y., Lin, D.: Spatial temporal graph convolutional networks for skeleton-based action recognition. In: AAAI, vol. 32 (2018)
35. Yang, S., Liu, J., Lu, S., Er, M.H., Kot, A.C.: Skeleton cloud colorization for unsupervised 3d action representation learning. In: ICCV, pp. 13423–13433 (2021)
36. Yue, R., Tian, Z., Du, S.: Action recognition based on RGB and skeleton data sets: a survey. Neurocomputing **512**, 287–306 (2022)
37. Zheng, N., Wen, J., Liu, R., Long, L., Dai, J., Gong, Z.: Unsupervised representation learning with long-term dynamics for skeleton based action recognition. In: AAAI, vol. 32 (2018)

Project and Pool: An Action Localization Network for Localizing Actions in Untrimmed Videos

Himanshu Singh[1], Avijit Dey[2], Badri Narayan Subudhi[1(✉)], and Vinit Jakhetiya[1]

[1] Indian Institute of Technology Jammu, Jammu and Kashmir, India
subudhi.badri@gmail.com

[2] Indian Association for the Cultivation of Science, Jadavpur, WestBengal, India

Abstract. Action recognition and localization in a video are challenging tasks in video analysis, requiring detecting and localizing actions within video sequences. Recent research has increasingly focused on enhancing the modeling of long-term temporal context. To address the said task in this paper, we have proposed a novel project and pool architecture. The proposed architecture comprises of three modules. In the initial module, we proposed LSTMProjector, which is a two-layer long-short-term memory module that projects spatial and temporal features from extracted videos in feature space. It efficiently handles input features by leveraging both local and global context information. The first LSTM layer processes each feature channel independently to capture local spatial dependencies, while the second layer captures global temporal dependencies across the entire sequence. In the second module, we devise a latent space projection technique to project the extracted features into a latent space using a one-dimensional convolutional layer to match the dimensions of spatial and temporal features. In the final module, a temporal pooling module is designed, which is a parameter-free max-pooling block and operates on local regions. It enhances the efficiency of the action localization model by selectively extracting the most crucial information from neighbouring and local clip embedding. We have demonstrated the effectiveness of the proposed scheme, using mean average precision (mAP) over different thresholds of Intersection over Union (IoU) on the "THUMOS 14", "Epic-Kitchens", and "MultiTHUMOS" datasets. The proposed technique achieves a mean average precision (mAP) of 67.38% on THUMOS 14, 24.71% on Epic-Kitchens verb, 23.30% on Epic-Kitchens noun, and 29.91% on MultiTHUMOS datasets. Additionally, we compared the performance of the proposed techniques with those of the twenty-eight state-of-the-art (SOTA) techniques on different benchmark databases, confirming the superiority of our proposed scheme.

Keywords: Action localization · Long short term memory · convolutional neural network

1 Introduction

Action recognition and localization are crucial tasks in computer vision. It has numerous applications, including human-computer interaction [29], monitoring kids/old-aged

A. Antonacopoulos et al. (Eds.): ICPR 2024, LNCS 15329, pp. 164–178, 2025.
https://doi.org/10.1007/978-3-031-78110-0_11

in smart homes [32], smart surveillance systems [14,36] and night surveillance systems [35], etc. Action localization specifically focuses on identifying an action's start and end times in an untrimmed video, while recognition concentrates on recognizing the category of each action instance. Trimmed videos, which contain only one action throughout, are temporally aligned and easier to analyze. On the other hand, untrimmed videos include additional irrelevant information, which makes it difficult to classify actions present in the video. In real-time applications, most videos are untrimmed and vary in duration. Typically, only a small portion of an untrimmed video contains relevant actions. So one approach to handle untrimmed videos is to crop or segment longer videos into smaller inputs before processing. However, the said methods risk missing parts of the action or failing to capture the entire action sequence. So, another optimal solution is to input the entire video into the network. Yet, due to computational limitations, processing full-length videos of varying durations remains challenging.

Several pioneer state-of-the-art (SOTA) [2,27,46] techniques generate classification score sequences over time by sliding a trained model across an untrimmed video or incorporating modules that capture long-range temporal relationships present between the video frames [25,30,43,49]. Analysis of various SOTA techniques reveals that recognizing and localizing actions in long untrimmed videos remains challenging, as these methods are unable to fully utilize the long-term temporal dependencies present in the videos. Furthermore, the complexity of actions, which often consist of multiple subactions, poses a challenge for models aiming to accurately identify the complete action. Further, current CNN-based SOTA techniques struggle to capture spatio-temporal features and are unable to extract relevant information from untrimmed videos. This highlights the importance of choosing a robust feature extraction method for action localization. Similarly, most CNNs and graph convolution networks are ineffective in capturing temporal correlations, particularly for long-term dependencies [15,30,44,50].

Based on the aforementioned analysis, we devise a novel action recognition and localization approach, "Project and Pool", from long untrimmed videos comprising three essential modules. Firstly, the LSTMProjector module is introduced, incorporating a two-layer long-short-term memory (LSTM) structure. The second module employs a latent space projection technique to transform the extracted features using a one-dimensional convolutional layer. Lastly, the third module introduces a temporal pooling mechanism, which employs a parameter-free max-pooling block that operates on local regions. The main contributions of this article are as follows:

- We have proposed an LSTM-based feature projection module, LSTMProjector, which brings out the inherent local and global details present between the video frames. This module effectively handles spatial and temporal features extracted from videos by integrating both local and global contextual information. The first LSTM layer works as an Intra-snippet feature projection layer to capture local spatial dependencies, while the second LSTM layer works as an Inter-snippet feature projection layer to capture long-range temporal dependencies within the entire video sequence.
- We have proposed a one-dimensional convolutional layer-based latent space module to enhance the LSTM projected features.

- A simple parameter-free temporal pooler module is introduced to exploit the temporal dependencies between the video frames.
- Our experiments on the THUMOS14, Epic-Kitchens and MultiTHUMOS datasets have been thorough, and the results show that our proposed method surpasses existing SOTA techniques for action localization.

We have utilized intersection over union (IoU) as our evaluation measure to show the effectiveness of the proposed scheme. The evaluation of the proposed technique is verified using twenty-three SOTA techniques on THUMOS14, three SOTA techniques for Epic-Kitchens and six SOTA techniques for MultiTHUMOS datasets to corroborate our findings. The organization of the remaining portion of this article is as follows: In Sect. 2, we have provided a brief discussion about the existing SOTA techniques. The proposed methodology is discussed in detail in Sect. 3. Section 4 represents experimental results with quantitative evaluations of the same. Finally, we conclude the proposed scheme in Sect. 5.

2 Related Work

Several human action recognition and localization algorithms have been developed during the last few decades, varying in technology/algorithm utilization, enhancement efficiency, and easy implementation. Despite advancements, action recognition and localization in untrimmed videos remain challenging. Several SOTA techniques like 3D-CNN [9] and I3D [4] are explored for action localization in untrimmed videos. These models are commonly utilized for feature extraction from videos, enabling subsequent application in action localization tasks. In the following section, we comprehensively analyzed other relevant SOTA techniques by classifying them into three specific types as follows:

One-Stage Action Localization. In the single-stage method, action localization is done without generating action proposals. Recently, action localization models have gained more attention in the computer vision community. These solutions aim to rapidly localise actions in the video without requiring separate action suggestions. The single-stage action localization technique often relies on anchor-based methods, in which anchors (windows) are extracted from the video using a sliding window method. Inspired by the Single-shot multi-box detector [24], Lin *et al.* [19] proposed the first single-stage, single-shot temporal action detection network. Buch *et al.* [3] proposed a single-stage recurrent memory-based model for temporal action localization. Based on a one dimensional convolutional network, Long *et al.* [27] used Gaussian kernels to optimise each anchor's size. Using convolutional networks again, Yang *et al.* [45] investigated the combination of anchor-based and anchor-free models for single-stage action localization. More recently, Lin *et al.* [18] created a saliency-based refinement module that was integrated into a convolutional network to propose an anchor-free single-stage model.

Two-Stage Action Localization. In this method, candidate video segments are first generated as action proposals; these proposals are further classified into different actions. After that, temporal boundaries are refined. Action proposals can be generated by using different ways: classification of anchor window [8,13], detecting action boundaries [11,20,21,52]. Some of the SOTA techniques have merged classification and proposal generation into a single model [5,34]. More recent SOTA techniques focus on modelling temporal context among action suggestions using an image-based visual-language model [16] or graph neural networks [44,48,50].

Transformer-Based Methods. Inspired by the great success of the Transformer in the field of machine translation and object detection, some recent works [12,25] adopt the attention mechanism in action localization tasks to improve the performance. Recent works [39,47] utilized the DEtection TRansformer-based decoder [54], which models action instances as a set of learnable parameters. On the other hand, Zhang *et al.* [49] encodes the formerly extracted features using a transformer-based encoder. However, most of these methods are based on local behaviour. Generally, attention operations are applied only within a local window, introducing an inductive bias similar to that of CNNs. However, this approach comes with increased computational complexity and additional limitations.

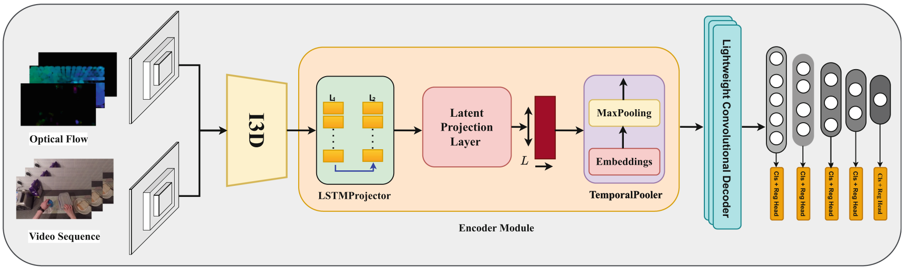


Fig. 1. Proposed framework for action localization.

3 Proposed Techniques

Figure 1 illustrates the framework of our proposed action localization technique. As shown in Fig. 1, we initially extracted the spatio-temporal features from the input videos using an Inflated 3D convolutional neural network [4]. These features are then fed into a LSTMProjection module followed by a latent space projection. The output from the latent space projection is then processed by a TemporalPooler module. After that, a lightweight convolutional decoder is applied. Finally, a classification and regression head is applied to generate the final output. We formulate the problem statement as follows:

3.1 Problem Statement

Given an untrimmed video X= $\{x_1, x_2, \ldots, x_T\}$ that contains "T" numbers of frames, our model seeks to predict a set of action instances, $\hat{\Phi} = \{\hat{\phi}_1, \hat{\phi}_2, \ldots, \hat{\phi}_M\}$. Here, M represents the number of action instances in X. Each action instance $\hat{\phi}_i$ = (s_t, e_t, a_c) consist of a start time (s_t), end time (e_t) along with its action label (a_c). The constraints are that s_t and e_t must be within the range $[1, T]$, $s_t < e_t$ and the action label a_c must belong to a predefined set of C action classes.

To solve the action localization problem, we initially devised a novel LSTM-based feature projection module to capture temporal dependencies present with video frames. We first describe the representation of the actions and then explain the LSTM-based encoder as follows;

Representation of Actions. In the proposed method, we adopt an anchor-free single-stage representation for each action occurrence [18,49]. This approach involves simultaneously regressing the onset and offset of an action based on the current time step, while the classification distinguishes between background and one of the C action classes at every time step. The method formulates the prediction for action localization as a sequence labeling task, which maps the video sequence X to a sequence of predicted action instances $\hat{\Phi}$ as follows:

$$X = \{x_1, x_2, ..., x_T\} \rightarrow \hat{\Phi} = \{\hat{\phi}_1, \hat{\phi}_2, ..., \hat{\phi}_M\} \tag{1}$$

At time step t, the output is given as $\hat{\phi}_t = (s^o_t, e^o_t, p^o_t)$. Where $s^o_t > 0$ and $e^o_t > 0$ represent the starting and end time of temporal spans of a given event. Considering C is the total number of action categories, the action probability p^o_t can be seen as a collection of $p^o_{j,t}$ denoting the likelihood for the j^{th} action at t time-step, where $1 \leq j \leq C$.

The predicated action instance p_{pt} at time step t can be inferred from $\hat{\phi}_t = (s^o_t, e^o_t, p^o_t)$ by:

$$p_{pt} = \arg\max(p^o_t), \quad s_t = t - s^o_t, \quad e_t = t + e^o_t \tag{2}$$

where s_t and e_t are the final predicted start and end times, respectively. The proposed method learns to label every input of X by $G(X) \rightarrow \hat{\Phi}$. Where G represents a deep learning model. G contains an encoder and decoder-like architecture $E \circ D$, which is given as follows;

3.2 Encoder Module

In the encoder module, we first represent the input feature X by a multi-scale temporal feature pyramid $Y = \{Y_1, Y_2, ..., Y_L\}$. The encoder contains the LSTM projection module followed by a one-dimensional convolution layer and a temporal pooler module at the end.

LSTM Projector. As shown in Fig. 2, in the LSTM Projector framework, firstly, the extracted features are reshaped and passed to the initial layer to capture temporal dependencies present within the feature sequence. The output is then reshaped and passed to the subsequent layer to capture temporal dependencies within the feature channels. The encoder E simply contains LSTMProjector as the feature projection layer, followed by $L - 1$ temporal context modeling blocks to produce feature pyramid Y. Formally, the LSTM-based projection layers are defined as:

$$X_P^1 = \text{LSTMProjector}(\text{Concat}(X)) \tag{3}$$

We first concatenate the input $X = \{x_1, x_2, ..., x_T\}$, in the channel dimension, then fed it into the LSTM projection module as described in equation (3). After that, it is passed to a latent projection layer, which is a one-dimensional convolutional layer resulting in $X_P^2 \in \mathbb{R}^{T \times D}$, which is given as follows:

$$X_P^2 = \text{Conv1D}(X_P^1) \tag{4}$$

This module's primary goal is to extract complex temporal patterns from extracted features. Each obtained feature encapsulates the essence of a video clip at a specific moment, capturing spatial and temporal information relevant to localizing actions in a video.

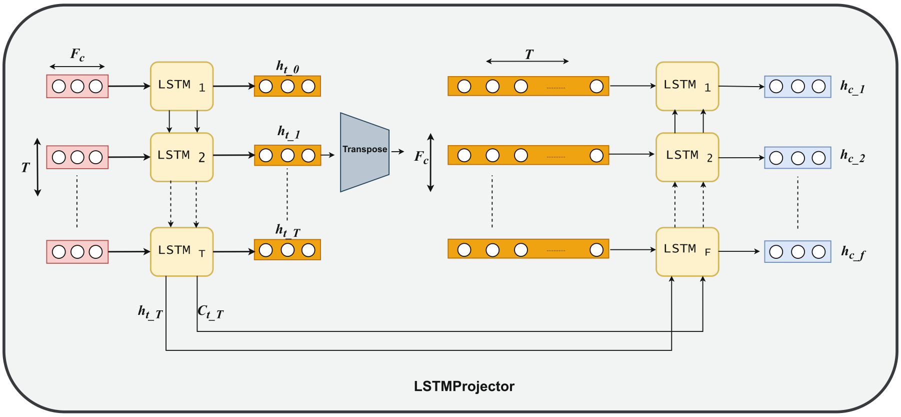


Fig. 2. Graphical representation of the LSTM Projector

The input tensor X is reshaped as X' and visualized as follows:

$$X' = \left[\begin{bmatrix} x_{1,1} \\ \vdots \\ x_{F,1} \end{bmatrix}, \begin{bmatrix} x_{1,2} \\ \vdots \\ x_{F,2} \end{bmatrix}, \ddots, \begin{bmatrix} x_{1,T} \\ \vdots \\ x_{F,T} \end{bmatrix} \right] \tag{5}$$

Where F represents the number of features, and T represents the number of sequences. The dimension of obtained $X^{'}$ is given as $(B \times C_h, T, 1)$. Where B represents the batch size, C_h represents the number of channels, and T represents the total time steps. We pass each feature of $X^{'}$ to an LSTM projection module individually. This step aims to capture temporal dependencies within each feature sequence. For each feature of $X^{'}$, LSTM projections are applied sequentially as follows:

$$\begin{aligned}
&\text{For } X_{i1}:\\
&h_{i1}^{(1)}, C_{i1}^{(1)} = LSTM_1^{FC}(x_1, h_{i0}, C_{i0})\\
&\text{For } X_{i2}:\\
&h_{i2}^{(2)}, C_{i2}^{(2)} = LSTM_2^{FC}(x_2, h_{i1}^{(1)}, C_{i1}^{(1)})\\
&\vdots\\
&\text{For } X_{iN}:\\
&h_{iT}^{(T)}, C_{iT}^{(T)} = LSTM_T^{FC}(x_T, h_{iT-1}^{(1)}, C_{iT-1}^{(1)})
\end{aligned}$$

Where, i = $\{1, 2, ..., F\}$ and $LSTM^{FC}$ represents the lstm layer applied along the feature channel.

After obtaining the hidden states for each feature X^i, we have h_i with shape $(B \times C, T, 1)$. Each h_i represents the hidden states obtained from i^{th} $LSTM$ projections for each feature. We concatenate all hidden states along the channel dimension to form h with shape $(B \times C, T, 1)$.The concatenated tensor h is visualized as:

$$h = \begin{bmatrix} \begin{bmatrix} h_{1,1}^{(1)} \\ \vdots \\ h_{F,1}^{(1)} \end{bmatrix}, \begin{bmatrix} h_{1,2}^{(2)} \\ \vdots \\ h_{F,2}^{(2)} \end{bmatrix}, \therefore, \begin{bmatrix} h_{1,T}^{(T)} \\ \vdots \\ h_{F,T}^{(T)} \end{bmatrix} \end{bmatrix} \tag{6}$$

we reshape the hidden state tensor h from $(B \times C, T, 1)$ to $(T, B \times C, 1)$ by permuting the dimensions.The reshaped hidden state tensor h^T is visualized as:

$$h^T = \begin{bmatrix} \begin{bmatrix} h_{1,1}^{(1)} \; h_{1,2}^{(2)} \cdots h_{1,T}^{(T)} \end{bmatrix} \\ \\ \vdots \\ \begin{bmatrix} h_{F,1}^{(1)} \; h_{F,2}^{(2)} \cdots h_{F,T}^{(T)} \end{bmatrix} \end{bmatrix} \tag{7}$$

After reshaping the hidden state tensor h, we pass each row of the reshaped tensor to the LSTM projections individually. Each row represents a separate sequence of hidden states across all time steps. LSTM projections are applied to each row of h^T as follows:

For the 1-st row of h^T :

$$h_1, C_1 = LSTM_1^{Temporal}([h_{1,1}^{(1)}, h_{1,2}^{(2)}, \ldots, h_{1,T}^{(T)}], h_{iT}^{(1)}, C_{iT}^{(1)})$$

$$\vdots$$

For the F-th row of h^T :

$$h_F, C_F = LSTM_F^{Temporal}([h_{F,1}^{(1)}, h_{F,3}^{(2)}, \ldots, h_{F,T}^{(T)}], h_{F-1}, C_{F-1})$$

Here, $LSTM^{Temporal}$ represents the LSTM layer applied along the temporal dimension.

Finally output hidden state (h_{out}) is given by;

$$h_{out} = \begin{bmatrix} h_1 = \begin{bmatrix} h_{1,1}^{(1)} & h_{1,2}^{(2)} & \cdots & h_{1,T}^{(T)} \end{bmatrix} \\ \vdots \\ h_F = \begin{bmatrix} h_{F,1}^{(1)} & h_{F,2}^{(2)} & \cdots & h_{F,T}^{(T)} \end{bmatrix} \end{bmatrix} \tag{8}$$

We reshape the concatenated $LSTM^{Temporal}$ outputs to match the original tensor dimension. Finally, a mask to the output tensor is applied.

TemporalPooler. We introduced a parameter-free max-pooling on the latent projected layer to emphasize the most salient features. This method captures temporal context by downsampling the input sequence while retaining the most relevant elements. This approach offers several advantages. It allows the model to capture task-specific, high-impact temporal features, adapting to diverse temporal dynamics. This adaptability enhances the model's capability to effectively interpret and analyze the varying complexities in different video sequences, leading to a more accurate understanding of complex patterns within the videos.

3.3 Decoder Design

The decoder D learns to predict sequence labeling, $\hat{\Phi} = \{\hat{\phi}_1, \hat{\phi}_2, ..., \hat{\phi}_M\}$, for every moment using multi-scale feature pyramid $Y = \{Y_1, Y_2, ..., Y_L\}$. The decoder adopts a lightweight convolutional neural network and consists of classification and regression heads. Formally, the two heads are defined as:

$$C_L = F_c(X_P^4(X_P^3(Y_L))) \tag{9}$$

$$O_L = \text{ReLU}(F_o(X_P^6(X_P^5(Y_L)))) \tag{10}$$

Here, $Y_L \in \mathbb{R}^{(T/2^{l-1}) \times C}$ is the latent feature of level L. X_P^3, X_P^4, X_P^5 and X_P^6 are the latent projections applied on each level of the feature pyramid Y_L, C_L represents the classification probability, and O_L represents the onset and offset prediction of the input moment set. The decoder architecture leverages the multi-scale features in the pyramid to make predictions for action localization.

3.4 Loss Function

We have utilized two loss functions, focal loss [22] and Distance-IoU loss [53]. The focal loss is used for binary classification, while Distance-IoU is used for distance regression. The overall loss for the proposed model is given as follows:

$$L_{total} = \frac{1}{T^{+}} \sum_{t} (L_{cls} + \lambda_{reg} \cdot 1_{c_t} \cdot L_{reg}) \tag{11}$$

where L_{cls} and L_{reg} represent the classification and regression loss, respectively. T^{+} is the total number of positive samples, and 1_{c_t} is an indicator function that denotes if a time step t is within an action, i.e., a positive sample. λ_{reg} represents a balancing coefficient for classification and regression loss. We have empirically set the value of λ_{reg} to 1. L_{total} is applied to all levels on the output pyramid and averaged across all video samples during training. Importantly, L_{cls} uses Focal loss to recognize all action classes. Focal loss naturally handles imbalanced samples. L_{reg} is only enabled when the current time step contains a positive sample.

4 Experimental Results

The proposed architecture is implemented using the open-source machine learning framework with PyTorch [38] on an Intel Xeon(R) Silver 4309Y CPU @ 2.80GHz system with 256GB RAM and NVIDIA A10080GB GPU.

4.1 Datasets

We have conducted experiments using the proposed technique for action localization on three benchmark datasets: THUMOS14, EPIC-Kitchens and MultiTHUMOS datasets. These datasets are widely used in the field of action localization and provide diverse and challenging video sequences for evaluation. By applying our technique to these datasets, we aimed to evaluate their performance in accurately localizing actions within videos and classifying them into respective action categories. THUMOS14, EPIC-Kitchens and MultiTHUMOS datasets offer a comprehensive range of action classes, temporal annotations, and many video samples, making them suitable for evaluating action localization and classification methods.

4.2 Evaluation Metric

We use the mean average precision (mAP), which is computed at different temporal intersections over union (tIoU). The 1D Jaccard index, or tIoU, is the intersection over union of two temporal windows. Based on the specified tIoU thresholds, we present the mAP scores for each action category. In addition, we report the average mAP value for all tIoU thresholds. The Intersection-over-union (IoU) measure can be given as

$$IoU = \frac{\text{Predicted action class} \cap \text{Ground truth}}{\text{Predicted action class} \cup \text{Ground truth}}.$$

The Mean average precision (mAP) is given by

$$mAP = \frac{TP}{TP + FP} = \frac{\text{Count of proposals predicted correctly}}{\text{Count of predicted proposals in total}}$$

4.3 Implementation Details

We have adhered to 3D-CNN to extract features from the video. First, we have divided the untrimmed videos into the non-overlapping chunk of 16 frames. We refer to these chunks as snippets. Then, we utilized the I3D model [4], pre-trained on the Kinetics-400 video dataset to extract spatiotemporal features. The snippets of the untrimmed videos are passed as input to the feature extractor. It converts these snippets to a vector of dimension 2048. which is then fed to the LSTM projection module, which provides an output of dimension 2034×2048. After that, one-dimensional convolution is applied, which provides a feature of dimension 2034×512. This is further processed by the temporal pooler module iteratively, generating a feature pyramid with five levels. The dimensions of each level are as follows: 2304×512, 1152×512, 576×512, 288×512, and 144×512. Each level has a classification and regression head to classify and localize the action. The ADAMW [28] optimizer with a learning rate of 0.0001 is utilized for training.

4.4 Experimental Results and Validation of Model

We have experimented with the proposed technique for action localization and classification on the Thumos 14, Epic-Kitchens and MultiTHUMOS datasets. Table 1 shows the performance comparison of the proposed method to twenty-three SOTA techniques for action localization for the THUMOS 14 dataset. The proposed scheme surpassed many SOTA techniques including the TadTR [25], AFSD [18], Re^2TL [51] and ActionFormer [49] by a margin of 20.78%, 15.38%, 2.48% and 0.58% in terms of average mAP respectively. Table 2 showcases performance comparisons for the Epic-Kitchens dataset for verb and noun tasks. It may be observed that the proposed method outperforms the other three SOTA approaches. Regarding average mAP, our suggested method surpasses BMN [20], G-TAD [44] and ActionFormer [49] by a margin of 16.31%, 15.31%, 1.21% for verb and 16.80%, 14.90%, 1.40% for noun task respectively. As shown in Table 3, Project and Pool showcase outstanding performance on the MultiTHUMOS dataset, achieving an average mean Average Precision (mAP) of 29.91%. Notably, our approach outshines the robust baseline, PointTAD [38], ASL [33] and ActionFormer [49], by a substantial margin. Furthermore, Project and Pool surpasses other long-term Temporal Context Modeling (TCM) methods, including those utilizing self-attention or Graph convolutional networks.

Table 1. Comparison with state-of-the-art methods on THUMOS 14 dataset

Type	Model	Publication	Feature	IoU					
				0.3	0.4	0.5	0.6	0.7	Avg.
Two-Stage	BMN [20]	ICCV2019	TSN [42]	56.0	47.4	38.8	29.7	20.5	38.5
	G-TAD [44]	CVPR 2020	TSN [42]	54.5	47.6	40.3	30.8	23.4	39.3
	DBG [17]	AAAI 2020	TSN [42]	57.8	49.4	39.8	30.2	21.7	39.8
	BC-GNN [2]	ECCV 2020	TSN [42]	57.1	49.1	40.4	31.2	23.1	40.2
	TAL-MR [52]	ECCV 2020	I3D [4]	53.9	50.7	45.4	38.0	28.5	43.3
	TCANet [30]	CVPR 2021	TSN [42]	60.6	53.2	44.6	36.8	26.7	44.3
	TSA-Net [11]	ICME 2020	P3D [31]	61.2	55.9	46.9	36.1	25.2	45.1
	P-GCN [48]	ICCV 2019	I3D [4]	63.6	57.8	49.1	—	—	—
	BMN-CSA [37]	ICCV 2021	TSN [42]	64.4	58.0	49.2	38.2	27.8	47.7
	RTD-Net [39]	ICCV 2021	I3D [4]	68.3	62.3	51.9	38.8	23.7	49.0
	VSGN [50]	ICCV 2021	TSN [42]	66.7	60.4	52.4	41.0	30.4	50.2
	P-GCN [48]+TSP [1]	ICCV 2019	R(2+1)D [41]	69.1	63.3	53.5	40.4	26.0	50.5
	ContextLoc [56]	ICCV 2021	I3D [4]	68.3	63.8	54.3	41.8	26.2	50.9
	MUSES [26]	CVPR 2021	I3D [4]	68.9	64.0	56.9	46.3	31.0	—
Single-Stage	A^{2}Net [45]	TIP 2020	I3D [4]	58.6	54.1	45.5	32.5	17.2	41.6
	GTAN [27]	CVPR 2019	P3D [31]	57.8	47.2	38.8	—	—	—
	PBRNet [23]	AAAI 2020	I3D [4]	58.5	54.6	51.3	41.8	29.5	—
	TadTR [25]	TIP 2022	I3D [4]	62.4	57.4	49.2	37.8	26.3	46.6
	AFSD [18]	CVPR 2021	I3D [4]	67.3	62.4	55.5	43.7	31.1	52.0
	ActionFormer [49]+TSP [1]	ECCV 2022	R(2+1)D [41]	73.4	67.4	59.1	46.7	31.5	55.6
	ContextLoc++ [55]	PAMI 2023	I3D [4]	74.4	68.2	58.7	46.3	30.8	55.7
	Re^{2}TL [51]	ICCV 2023	SlowFast [10]	77.4	72.6	72.4	53.7	39.0	64.9
	ActionFormer [49]	ECCV 2022	I3D [4]	82.1	77.8	71.0	59.4	43.9	66.8
	Proposed Method	-	I3D	**83.15**	**78.82**	**71.07**	**58.59**	**44.41**	**67.38**

Table 2. Comparison with state-of-the-art methods on EPIC-Kitchens dataset

Task	Model	IoU					Avg.
		0.1	0.2	0.3	0.4	0.5	
Verb	BMN [20]	10.8	9.8	8.4	7.1	5.6	8.4
	G-TAD [44]	12.1	11.0	9.4	8.1	6.5	9.4
	ActionFormer [49]	26.6	25.4	24.2	22.3	19.1	23.5
	Proposed Method	**27.7**	**26.75**	**25.63**	**23.25**	**20.22**	**24.71**
Noun	BMN [20]	10.3	8.3	6.2	4.5	3.4	6.5
	G-TAD [44]	11.0	10.0	8.6	7.0	5.4	8.4
	ActionFormer [49]	25.2	24.1	22.7	20.5	17.0	21.9
	Proposed Method	**26.94**	**25.92**	**24.14**	**21.55**	**17.97**	**23.30**

Table 3. Comparison with the state-of-the-art methods on the MultiTHUMOS dataset. We report the results at different IoU thresholds [0.2, 0.5, 0.7] and average mAP in [0.1:0.9:0.1]

Method	IoU			
	0.2	0.5	0.7	Avg.
MLAD [40]	—	—	—	14.2
MS-TCT [6]	—	—	—	16.2
PDAN [7]	—	—	—	17.3
PointTAD [38]	39.7	24.9	12.0	23.5
ASL [33]	42.4	27.8	13.7	25.5
ActionFormer [49]	46.4	32.4	15.0	28.6
Proposed Method	**46.95**	**33.42**	**17.97**	**29.91**

5 Conclusion

This paper addresses the important task of action localization in long, untrimmed videos. The proposed method introduces several key components to improve localization accuracy. First, we have devised a novel LSTM projection module. Then, we have further projected them into a latent space using a one-dimensional convolutional layer. Finally, we integrated a temporal pooling module, a straightforward, parameter-free max-pooling block that functions on the local region. This module captures long-range spatial-temporal dependencies. A combination of focal loss and Distance-IoU loss functions is employed to train the network. To evaluate the efficiency of the proposed scheme, experiments were conducted on the THUMOS 14, Epic-Kitchens and MultiTHUMOS video datasets. The performance of the proposed approach was compared against twenty-three SOTA techniques on the THUMOS 14, three SOTA techniques on the Epic-Kitchens and six SOTA techniques on the MultiTHUMOS datasets. The evaluation metrics used include intersection over union (IoU). The findings of the experiments support the effectiveness of the proposed scheme in action localization, as it outperformed the compared SOTA techniques on all datasets. The use of an LSTM projector and temporal pooler contributes to the improved accuracy and efficiency of the proposed approach.

References

1. Alwassel, H., Giancola, S., Ghanem, B.: TSP: temporally-sensitive pretraining of video encoders for localization tasks. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 3173–3183 (2021)
2. Bai, Y., et al.: Boundary content graph neural network for temporal action proposal generation. In: Proceedings of the European Conference on Computer Vision, pp. 121–137. Springer, Cham (2020)
3. Buch, S., et al.: SST: single-stream temporal action proposals. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2911–2920 (2017)
4. Carreira, J., Zisserman, A.: Quo vadis, action recognition? A new model and the kinetics dataset. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 6299–6308 (2017)

5. Chao, Y.-W., et al.: Rethinking the faster R-CNN architecture for temporal action localization. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1130–1139 (2018)
6. Dai, R., et al.: MS-TCT: multi-scale temporal convtransformer for action detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 20041–20051 (2022)
7. Dai, R., et al.: PDAN: pyramid dilated attention network for action detection. In: Proceedings of the IEEE Winter Conference on Applications of Computer Vision, pp. 2970–2979 (2021)
8. Escorcia, V., et al.: DAPS: deep action proposals for action understanding. In: Proceedings of the European Conference on Computer Vision, pp. 768–784 (2016)
9. Fan, H., et al.: Reconfigurable acceleration of 3D-CNNs for human action recognition with block floating-point representation. In: Proceedings of the International Conference on Field Programmable Logic and Applications, pp. 287–2877 (2018)
10. Feichtenhofer, C., et al.: Slowfast networks for video recognition. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 6202–6211 (2019)
11. Gong, G., Zheng, L., Mu, Y.: Scale matters: temporal scale aggregation network for precise action localization in untrimmed videos. In: Proceedings of the IEEE International Conference on Multimedia and Expo, pp. 1–6 (2020)
12. Gritsenko, A.A., et al.: End-to-end spatio-temporal action localisation with video transformers. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 18373–18383 (2024)
13. Heilbron, F.C., Niebles, J.C., Ghanem, B.: Fast temporal action formervity proposals for efficient detection of human actions in untrimmed videos. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1914–1923 (2016)
14. Kapoor, M., et al.: Underwater moving object detection using an end-to-end encoder-decoder architecture and GraphSage with aggregator and refactoring. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, pp. 5636–5645 (2023)
15. Kipf, T.N., Welling, M.: Semi-Supervised Classification with Graph Convolutional Networks. (2017). arXiv: 1609.02907
16. Li, Z., et al.: DeTAL: open-vocabulary temporal action localization with decoupled networks. IEEE Trans. Pattern Anal. Mach. Intell. 1–14 (2024)
17. Lin, C., et al.: Fast learning of temporal action proposal via dense boundary generator. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 34, pp. 11499–11506 (2020)
18. Lin, C., et al.: Learning salient boundary feature for anchor-free temporal action localization. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3320–3329 (2021)
19. Lin, T., Zhao, X., Shou, Z.: Single shot temporal action detection. In: Proceedings of the ACM International Conference on Multimedia, pp. 988–996 (2017)
20. Lin, T., et al.: BMN: boundary-matching network for temporal action proposal generation. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 3889–3898 (2019)
21. Lin, T., et al.: BSN: boundary sensitive network for temporal action proposal generation. In: Proceedings of the European Conference on Computer Vision, pp. 3–19 (2018)
22. Lin, T.-Y., et al.: Focal loss for dense object detection. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 2980–2988 (2017)
23. Liu, Q., Wang, Z.: Progressive boundary refinement network for temporal action detection. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 34, pp. 11612–11619 (2020)
24. Liu, W., et al.: SSD: single shot multibox detector. In: Proceedings of the European Conference on Computer Vision, pp. 21–37 (2016)

25. Liu, X., et al.: End-to-end temporal action detection with transformer. IEEE Trans. Image Process. **31**, 5427–5441 (2022)
26. Liu, X., et al.: Multi-shot temporal event localization: a benchmark. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 12596–12606 (2021)
27. Long, F., et al.: Gaussian temporal awareness networks for action localization. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 344–353 (2019)
28. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. In: Proceedings of the International Conference on Learning Representations, pp. 1–19 (2018)
29. Meng, H., Pears, N., Bailey, C.: A human action recognition system for embedded computer vision application. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1–6 (2007)
30. Qing, Z., et al.: Temporal context aggregation network for temporal action proposal refinement. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 485–494 (2021)
31. Qiu, Z., Yao, T., Mei, T.: Learning spatio-temporal representation with pseudo-3D residual networks. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 5533–5541 (2017)
32. Rougier, C., et al.: Robust video surveillance for fall detection based on human shape deformation. IEEE Trans. Circuits Syst. Video Technol. **21**(5), 611–622 (2011)
33. Shao, J., et al.: Action sensitivity learning for temporal action localization. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 13457–13469 (2023)
34. Shou, Z., et al.: CDC: convolutional-de-convolutional networks for precise temporal action localization in untrimmed videos. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 5734–5743 (2017)
35. Singh, H., et al.: Action recognition in dark videos using spatio-temporal features and bidirectional encoder representations from transformers. IEEE Trans. Artif. Intell. **1**(1), 1–11 (2022)
36. Singh, H., et al.: C3D and localization model for locating and recognizing the actions from untrimmed videos (student abstract). In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 36, pp. 13051–13052 (2022)
37. Sridhar, D., et al.: Class semantics-based attention for action detection. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 13739–13748 (2021)
38. Tan, J., et al.: PointTAD: Multi-Label Temporal Action Detection with Learnable Query Points. arXiv preprint arXiv:2210.11035 (2022)
39. Tan, J., et al.: Relaxed transformer decoders for direct action proposal generation. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 13526–13535 (2021)
40. Tirupattur, P., et al.: Modeling multi-label action dependencies for temporal action localization. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1460–1470 (2021)
41. Tran, D., et al.: A closer look at spatiotemporal convolutions for action recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 6450–6459 (2018)
42. Wang, L., et al.: Temporal segment networks: towards good practices for deep action recognition. In: Proceedings of the European Conference on Computer Vision, pp. 20–36 (2016)
43. Wang, L., et al.: Temporal Action Proposal Generation with Transformers (2021). arXiv: 2105.12043
44. Xu, M., et al.: G-TAD: sub-graph localization for temporal action detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 10156–10165 (2020)

45. Yang, L., et al.: Revisiting anchor mechanisms for temporal action localization. IEEE Trans. Image Process. **29**, 8535–8548 (2020)
46. Yang, Z., Qin, J., Huang, D.: ACGNET: action complement graph network for weakly-supervised temporal action localization. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 36, pp. 3090–3098 (2022)
47. Yuan, L., et al.: Tokens-to-token VIT: training vision transformers from scratch on imagenet. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 558–567 (2021)
48. Zeng, R., et al.: Graph convolutional networks for temporal action localization. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 7094–7103 (2019)
49. Zhang, C.-L., Wu, J., Li, Y.: Actionformer: localizing moments of actions with transformers. In: Proceedings of the European Conference on Computer Vision, pp. 492–510 (2022)
50. Zhao, C., Thabet, A.K., Ghanem, B.: Video self-stitching graph network for temporal action localization. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 13658–13667 (2021)
51. Zhao, C., et al.: Re2TAL: rewiring pretrained video backbones for reversible temporal action localization. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 10637–10647 (2023)
52. Zhao, P., et al.: Bottom-up temporal action localization with mutual regularization. In: Proceedings of the European Conference on Computer Vision, pp. 539–555 (2020)
53. Zheng, Z., et al.: Distance-IoU loss: faster and better learning for bounding box regression. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 34, pp. 12993–13000 (2020)
54. Zhu, X., et al.: Deformable DETR: Deformable Transformers for End-to-End Object Detection. arXiv preprint arXiv:2010.04159 (2020)
55. Zhu, Z., et al.: Contextloc++: a unified context model for temporal action localization. IEEE Trans. Pattern Anal. Mach. Intell. (2023)
56. Zhu, Z., et al.: Enriching local and global contexts for temporal action localization. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 13516–13525 (2021)

Multi-teacher Importance Preserving Knowledge Distillation for Early Violence Prediction

Suvramalya Basak, Aditya Vaishy, and Anjali Gautam(✉)

Computer Vision and Biometrics Lab, Indian Institute of Information Technology
Allahabad, Prayagraj, U.P., India
anjaligautam@iiita.ac.in

Abstract. The real-world applicability of automated violence recognition systems has drawn much attention from researchers. The current techniques for recognizing violence are centered on creating efficient models that can predict violent events quickly and accurately in real-time. However, early violence prediction, which is crucial for real-time systems, is not considered in these methods. In this paper, we present an early violence prediction method that can accurately predict violent activities from partially observed video frames. We propose a two-stream architecture which employs our proposed efficient **S**queeze-**E**xcitation **S**huffle**Net** (SESNet) model that effectively extracts spatial and temporal features. We leverage spatio-temporal and channel-wise squeeze and excitation to incorporate attention information into the ShuffleNet V2 architecture. To enable early violence recognition, we train our model in a teacher-student framework, where the teacher model trained on full-length videos distils privileged information to the student model, which has access to partial videos. For this purpose, we introduce a novel multi-teacher importance preservation learning methodology which can effectively distill important features from multiple teacher networks. We evaluate our approach on the challenging RWF-2000 public violence recognition dataset. Experimental results show that our teacher-student training framework performed well for early violence prediction. Additionally, our proposed model also outperforms several state-of-the-art violence recognition methods on full-length videos.

Keywords: Early Violence Prediction · Knowledge Distillation · Efficient Convolution Networks · Autonomous Surveillance System

1 Introduction

The widespread deployment of surveillance cameras in public spaces has demonstrably yielded a safer environment. Their presence deters criminal activity and provides crucial evidence for investigations. However, the ever-increasing number of cameras translates to a vast amount of video data, posing a significant

A. Antonacopoulos et al. (Eds.): ICPR 2024, LNCS 15329, pp. 179–195, 2025.
https://doi.org/10.1007/978-3-031-78110-0_12

challenge for traditional monitoring methods. Manually sifting through hours of footage is not only time-consuming and labor-intensive but also hinders the proactive potential of surveillance systems. This has fueled a surge in research on automated violence recognition systems. Violence recognition aims at recognizing violent or aggressive human actions such as fighting, rioting, vandalism, etc.

Several earlier works on violence recognition made use of hand-crafted feature descriptors capable of capturing violence motion in video data [3,11]. With the success of deep learning in computer vision based applications [1,2], recent works have largely focused on deep learning based solutions to improve recognition performance [6,9,12]. However, these works intend to recognize violence from full videos, thus making them post-incident analysis tools. As these methods are trained on full videos, they learn more prominent features that may not be present in the partial video. As a result these methods are often incapable of recognizing violent activities when partial videos are provided, as can be seen in Fig. 1. The ability to predict violence from partial videos is extremely important for real-time violence recognition systems, as this allows authorities to receive accurate contemporary predictions rather than post-incident recognition. For example, in a real-world surveillance system, recognizing any violent behaviour after it has already occurred is not very meaningful. It would be more helpful if the system could predict an ongoing violence act as early as possible. This could help in prompt response by the authorities to stop or avoid such situations.

Although multiple research works have been done regarding early action prediction [28,30], hardly any focus is given to early violence prediction. We define early violence prediction as the task to accurately predict violent human behaviour from partially observed videos. To this end, we propose an early violence prediction model which is able to accurately predict violent activities from different lengths of partial videos. We first design our Efficient Violence Net (EVNet) architecture, which acts as the backbone for all our recognition tasks. Our EVNet is a two-stream architecture leveraging our proposed Squeeze-Excitation ShuffleNet (SESNet) model. Our SESNet presents an efficient 3D convolution architecture based on the ShuffleNetv2 model [17], enhanced by the spatio-temporal and channel squeeze and excitation. Further, to incorporate early violence detection capability, we train our EVNet architecture in a teacher-student setup using knowledge distillation. We also propose an importance preserving knowledge distillation loss for training our student model for early violence prediction.

We can sum up our main contributions as follows:

1. We propose a lightweight two-stream violence recognition model, called EVNet, which leverages our efficient SESNet architecture.
2. We train our EVNet model in a teacher-student setup to incorporate early violence prediction capability in our violence classifier. To the best of our knowledge, this is the first work to deal with early violence prediction.
3. An importance preserving multi-teacher knowledge distillation is proposed to further improve violence prediction performance from partially observed videos.

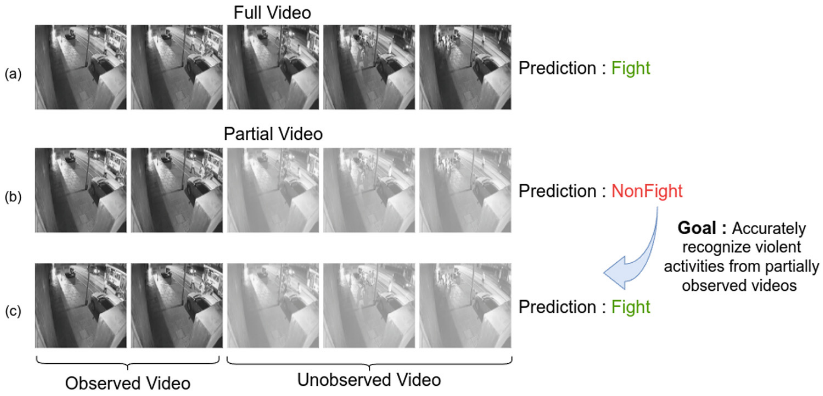


Fig. 1. Prediction of violence in the video (fight/no fight). Prediction of fight from the (a) full video, (b) partial video with few initial frames, however, model is giving wrong prediction, (c) the model is accurately predicting the violence in the video from few initial frames. Therefore, the task is to develop a model that can accurately predict violence from partially observed video.

The rest of this article is organised as follows. Relevant literature is summarised in Sect. 2. In Sect. 3, we detail our proposed EVNET and SESNet architectures for violence recognition, as well as our proposed multi-teacher important feature preserving distillation loss for early violence prediction. The experimental details and results are provided in Sect. 4 and Sect. 5 respectively. Section 6 concludes the paper.

2 Related Work

Traditional methods made use of feature descriptors to capture relevant spatio-temporal information from the videos. Nievas et al. [3] used the Motion Scale Invariant Feature Transform (MoSIFT) [5] and the Spatio-temporal Interest Points (STIP) [14] as feature descriptors, and finally classify violence videos utilizing a Bag-of-words framework. The work also introduced the popular Hockey Fights dataset. In [11], Hassner et al. proposed the Violent Flows (ViF) descriptor, which takes into consideration how the magnitude of optical flow vectors changes over time, instead of taking just the magnitude values themselves.

With the limited generalization ability of such methods, and the success of deep learning, various recent deep learning based methods have outperformed the traditional methods. Given their ability to extract spatio-temporal features, 3D Convolution Neural Networks (CNNs) have been extensively used in violence recognition [8,16]. Several works [6,7,9,12,20] in violence recognition adopted the two-stream approach proposed by [22]. Cheng et al. [6] introduced the RWF-2000 violence recognition dataset. Islam et al. [12] replace optical flows in [6] with RGB frame difference, which are faster to compute. Garcia et al. [9] replace RGB frames with skeleton features which are able to better capture motion for human objects. The use of Long Short-Term Memory (LSTM) networks have also been widely explored in violence recognition [7,12,25]. Dai et al. [7] applied LSTM over two stream architectures to enhance the capture of temporal features. Islam et al. [12] replaced ConvLSTMs with Separable Convolutional LSTM (SepConvLSTM) which further reduces model parameters. However, most of these work fail to incorporate attention mechanism when extracting features. Attention has proved to be extremely influential for extraction of important features and the omission of redundant ones [10]. In our work, we use a two-stream architecture where our squeeze-and-excitation attention equipped SESNet model extracts salient local spatio-temporal information from RGB and frame difference streams, and long range temporal dependencies are then learnt using temporal convolution network (TCN) [15].

Several works have focused on proposing efficient, lightweight models for violence recognition, which is beneficial in real-time setting. However, little focus is given to the early prediction of violence activities. Early prediction of violent activities are more valuable for authorities instead of post-event predictions. For general action anticipation task, Wang et al. [28] proposed a teacher student knowledge distillation approach where knowledge from a teacher model trained on full length videos are distilled into a student model which has access to partial videos. As there is a large information gap between the input to the teacher and student models, Zhao et al. [30] introduces a curriculum learning approach to distill knowledge from teacher model. Camporese et al. [4] approaches action anticipation as a multi-label task. The work looks to predict the action at a future time-step, using label-smoothing to remove uncertainty of future predictions. In this work, we put forth an early violence prediction model using importance preserving knowledge distillation. The issue with the knowledge distillation loss functions used in [28] is the loss of crucial information for individual progress levels, as the model tries to accommodate features for all progress levels. By distilling only the important features from our teacher models, we are able to preserve the information at every progress level. We detail our proposed approach in the next Sections.

3 Proposed Approach

The aim of this work is to design an early violence prediction model that is able to predict violent actions accurately from partially observed video samples. Given a training set $\mathcal{V} = \{V_i, y_i\}_{i=1}^{|\mathcal{V}|}$, where every video $V_i = \{v_j\}_{j=1}^{H}$ consists of H frames and has video-level label $y_i \in Y = \{0, 1\}$. Here, a video has $y_i = 1$ if it belongs to violence class, and $y_i = 0$ otherwise. $|\mathcal{V}|$ represents the number of videos in the training set. A partial video comprises of the first k frames of any given video V_i, and the observation ratio is given by k/H. The total number of progress levels are denoted by N.

For the purpose of early violence prediction, we design a teacher-student framework, the violence anticipation model, which has access to only partial videos, is trained with the help of a violence recognition teacher model trained

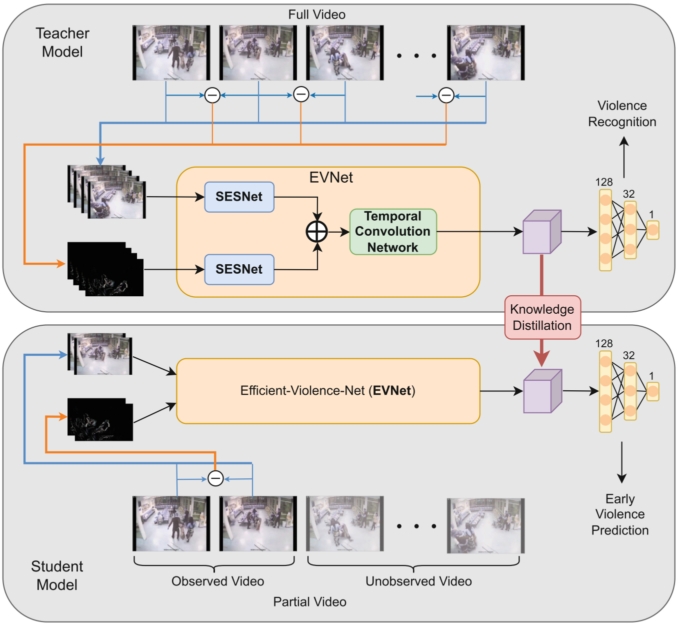


Fig. 2. Schematic diagram of our knowledge distillation based Early Violence Classifier. Our teacher EVNet is trained on full length videos for violence recognition task. The student model also has same EVNet architecture, however, it is trained on partially observed videos with knowledge distilled from important features in the teacher model. The trained student model is named as Efficient Early Violence Network (EEVNet).

on full length training videos. In this section, we first discuss the architecture of our violence recognition model, followed by the knowledge distillation approach used to train the early violence prediction model.

3.1 Squeeze-Excitation Shuffle Net Architecture

Given the application of violence anticipation in real-time violence prediction, it is important that our violence prediction model is lightweight and efficient. In this section, we introduce our **S**queeze-**E**xcitation **S**huffle **Net** (SESNet), inspired by [17] and [29] where we combined the concepts of shuffle attention [29] with the ShuffleNet V2 model [17] to form our SESNet. The structure of the building blocks of our SESNet are illustrated in Fig. 3. At the beginning of every block, the channel split operation is performed. This divides the input features of C channels into two parts with $\bar{C}$ and $C-\bar{C}$ channels, respectively. For this work, we set $\bar{C} = C/2$. For spatial down sampling, that is using $stride = 2$, we keep the architecture same as in the ShuffleNetV2 paper. However, for $stride = 1$, the last $C - \bar{C}$ channels are provided to our introduced attention module. The first $\bar{C}$ channels are passed onto the depthwise convolution layers, as in the original

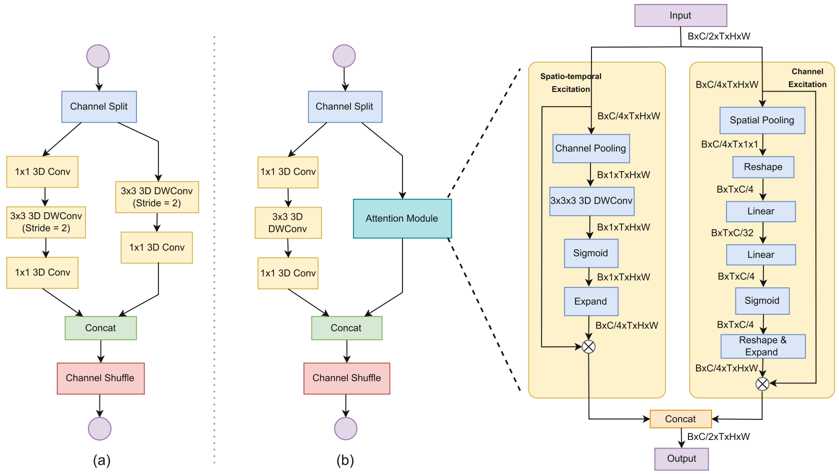


Fig. 3. Building blocks of our proposed SESNet. (a) Unit block for spatial down sampling (2x), similar to [17]. (b) Basic unit with squeeze and excitation attention module.

ShuffleNet V2 model. Finally, the outputs of the two branches are concatenated and provided to the Channel Shuffle block, same as in [17].

The attention module comprises of the channel excitation and spatio-temporal excitation modules. The input to the attention module is a tensor of shape $B \times C/2 \times T \times H \times W$, where B is the batch size, C the feature channels, and T, H, W are the sizes across temporal and spatial dimensions. The attention module further divides this tensor into two equal halves of shape $B \times C/4 \times T \times H \times W$, one is sent to the channel excitation module, whereas the other is sent to the spatio-temporal excitation module. The two excitation modules are discussed below.

Channel Excitation. The channel excitation (CE) module is designed to exploit channel dependencies. The global spatial information of the input is first squeezed with the help of a spatial average pooling layer.

The squeezed output is denoted by $X_p \in \mathbb{R}^{B \times C/4}$. We reshape X_p into $X_p^* \in \mathbb{R}^{B \times T \times C/4}$, which can now to supplied to the fully connected layer that squeeze the number of channels by ratio r. Another linear layer is used to restore the number of channels back to $C/4$. The sigmoid activation is then applied to the output from the linear layers. This is formulated as:

$$F_m = \sigma(\delta(X_p^* W_1) W_2) \tag{1}$$

where $F_m \in \mathbb{R}^{B \times C/4 \times T}$, δ denotes ReLU activation function, σ denotes sigmoid activation function, and W_1 and W_2 are the weights of the two linear layer, respectively. F_m is then expanded across spatial dimensions and reshaped to

$F_o \in \mathbb{R}^{B\times C/4\times T\times H\times W}$. The element-wise multiplication of F_o with input X is taken, and added to the input, which gives the final output. The output is formulated as:

$$O = F_o \otimes X \tag{2}$$

where $\otimes$ denotes the element-wise multiplication.

Spatio-Temporal Excitation. Similar to channel excitation, we use the spatio-temporal excitation (STE) to find spatial-temporal dependencies with the help to 3D convolution layer. First the input X is converted to $X_c \in \mathbb{R}^{B\times 1\times T\times H\times W}$ by a channel average pooling layer. This gives us the global spatio-temporal information. X_c is then provided to a depth-wise 3D convolution layer with $3 \times 3 \times 3$ kernel size. This is formulated as:

$$X_c^* = \mathcal{W} * X_c \tag{3}$$

where $\mathcal{W}$ denotes the kernel weights of the 3D convolution layer, and $*$ denotes the convolution operation. We pass the output to a sigmoid activation function which gives us the spatio-temporal mask $F_m = \sigma(X_c^*)$. The final output is given similar to Eq. 2.

The output of CE and STE modules are concatenated along the channel dimension to give the final attention module output $\mathcal{A}_o \in \mathbb{R}^{B\times C/2\times T\times H\times W}$.

3.2 Violence Recognition Model

The overall violence recognition model, called **E**fficient-**V**iolence-**Net** (EVNet), can be seen in Fig. 2. We use a two-stream approach. The first stream contains RGB frames, whereas the second stream contains frame difference information. Both inputs are provided to two separate SESNets. The output features of the two SESNets are then fused using element-wise addition, and then passed onto a Temporal Convolution Network (TCN) [15]. The TCN consists of one block with dilation factor of 2. Each TCN block contains two dilated causal convolution layers with kernel size of 5. The output of the TCN is provided to a fully connected classifier.

3.3 Early Violence Prediction Model

With the objective of allowing our model to perform early violence prediction, we train our model to classify violence from partially observed videos. In this section, we describe two different training strategies for achieving our objective.

Progressive Teacher-Student Learning. In this training method, we follow the progressive teacher-student method described in [28]. In detail, our VR model trained on full length videos acts as a teacher for the student network which is trained on partial videos of different progress levels. The student model has the same architecture as the teacher. We extract latent features from the TCN module from both the teacher and student models for all progress levels. We denote these feature representations as T_i and S_i for teacher and student respectively, where i denotes the $i-th$ video sample. Hence, T_i and S_i are two $D \times N$ feature vectors, D being the feature dimension. The student model is then trained using the loss formulated in [28], which is given as

$$L = \frac{1}{|\mathcal{V}|}\sum_{i=1}^{|\mathcal{V}|}(L_C(S_i, y_i) + L_{TS}(S_i, T_i)) \tag{4}$$

where y_i denotes the ground truth annotation for *i-th* video-sample and L_C is the standard binary cross entropy loss which acts as the classification loss of the student model. Here, L_{TS} is the knowledge distillation loss given as $L_{TS} = \alpha L_{MSE} + \beta L_{MMD}$, where L_{MSE} and L_{MMD} are the mean square error (MSE) and maximum mean discrepancy (MMD).

Multi-teacher Learning with Important Feature Preservation. As discussed in Sect. 1, the teacher model incorrectly classifies a partially observed video. Hence, distilling this information may not be beneficial in training the student model. To counter this issue, Zhao et al. [30] proposed a curriculum learning procedure using intermediate teachers. This requires the student model to be trained iteratively by all teachers, which is time consuming. In our work, we train specialized teachers for every progress level. These intermediate teachers are denoted as T_4, T_6, T_8 for models trained on 40%, 60% and 80% of observable full length videos respectively. The teacher trained on full videos is denoted as T_{10}. However, unlike [30], we train the student in one step, as described in [28]. For every progress level n, the student learns from knowledge distilled from the closest teacher model, as seen in Fig. 4. For example, for observation ratio 20%, knowledge from the teacher trained on 40% videos is distilled. Hence, for every progress level, the student learns from a specialized teacher. Let the set of teacher models be $\mathcal{T} \in [T_4, T_6, T_8, T_{10}, T_{10}]$, with latent features from T_{10} being used to train the student model for the last two progress levels. The final loss is formulated as,

$$L_{MT} = \frac{1}{|\mathcal{V}|}\sum_{i=1}^{|\mathcal{V}|}\frac{1}{|N|}\sum_{n=1}^{|N|} L_C(S_{n,i}, y_i) + \gamma L_{KD}(S_{n,i}, \mathcal{T}_{n,i}) \tag{5}$$

where $|N|$ is the number of teacher models in $\mathcal{T}$, L_C is binary cross entropy loss, and $L_{KD} = ||S_{n,i} - \mathcal{T}_{n,i}||_F^2$.

Training the student across all progress levels, the model learns features that are beneficial for all progress levels. However, to accommodate features of lower progress levels, the model loses crucial information for later progress levels. As a result, the student models performance for later progress levels are lacking when compared to the teacher model. This can be seen in Table 2 in Sect. 5. To alleviate this issue, we introduce important feature preserving knowledge distillation loss.

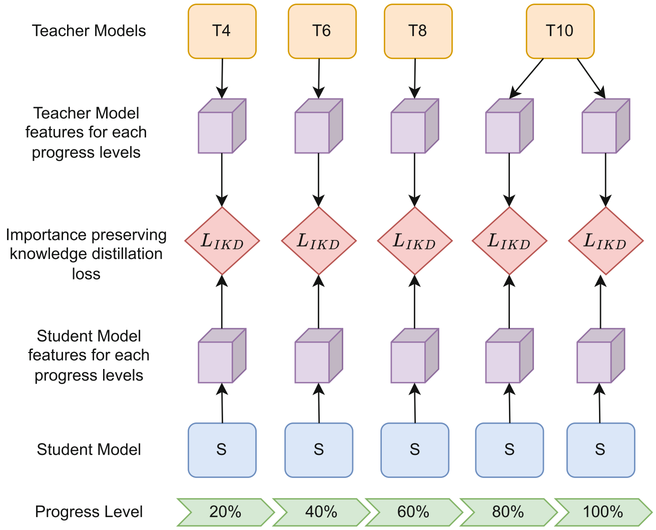

Teacher Models
T4
T6
T8
T10
Teacher Model features for each progress levels
Importance preserving knowledge distillation loss
L_{IKD}
Student Model features for each progress levels
Student Model
S
Progress Level
20%
40%
60%
80%
100%

Fig. 4. Schematic diagram of our Multi-Teacher learning approach.

Taking inspiration from incremental learning works [13,19], we distill knowledge from features that are important for the teacher model's performance. As a result, the student model also learns these important feature channels for each progress level, whereas less critical channels are allowed to be changed for other progress levels. The importance mask for every model $\mathcal{T}_n \in \mathcal{T}$ is calculated as follows

$$I_{\mathcal{T}_n} = \mathbb{E}_{(V_i,y_i)\sim\mathcal{V}}||\nabla_{F_{t,c}} L_{cls}(V_i, y_i)||_F^2 \tag{6}$$

where L_{cls} is the classification loss of the trained teacher model, V_i and y_i are the input video and its ground-truth label, $\mathbb{E}_{(V_i,y_i)\sim\mathcal{V}}$ is the expectation over all training videos in $\mathcal{V}$. Put simply, the channels in the feature map of the Frobenius norm of the gradient which, on perturbation, result in larger increase of final loss of the model, are considered to be important features. We normalize the importance mask $I_{\mathcal{T}_n}$ as

$$\bar{I_{\mathcal{T}_n}} = \frac{I_{\mathcal{T}_n}}{\frac{1}{T,C}\sum_{t=1}^{T}\sum_{c=1}^{C} I_{\mathcal{T}_n}} \tag{7}$$

Finally, we define our masked distillation loss as $L_{IKD} = \bar{I_{\mathcal{T}_n}}||S_{n,i} - \mathcal{T}_{n,i}||_F^2$. The overall loss is given as

$$L_{MTI} = \frac{1}{|\mathcal{V}|}\sum_{i=1}^{|\mathcal{V}|}\frac{1}{|N|}\sum_{n=1}^{|N|} L_C(S_{n,i}, y_i) + \gamma L_{IKD}(S_{n,i}, \mathcal{T}_{n,i}) \tag{8}$$

4 Experimental Settings

4.1 Datasets

Since goal objective is to recognize violent actions in surveillance scenario, we select the RWF-2000 dataset which is best suited for our task. The dataset contains 2000 surveillance videos, where 1000 depict fights, and the rest are normal activities. Further, the dataset is suitable for testing our early violence prediction model. We also perform early violence prediction tests on the Real-Life Violence Surveillance (RLVS) dataset [23]. The dataset contains 2000 videos of fights occuring in various settings.

4.2 Implementation Details

As violent activities are characterized by sharp and sudden motion. To extract minute motion information, we extract frames at frequent intervals from the videos. From each video, 75 frames are sampled and resized to 224 × 224. During training, the data is augmented using techniques such as color jitter, random cropping, Gaussian blurring, random flip to reduce overfitting. We use batch size of 8 and Adam optimizer with learning rate of 0.001. The values of α and β for L_{TS} loss in Eq. 4 are set to 0.1 and 0.002 respectively. The value of γ in Eq. 8 is set to 2. The SESNet contains 3 blocks with 2 layers in each block. Experiments have been performed on Nvidia RTX 3090 GPU and all codes have been implemented in Pytorch.

5 Results and Discussions

5.1 Results on Full Length Videos

In this section we compare the results of our models on the RWF-2000 dataset with existing methods. The results are summarized in Table 1. As is evident from the results, our violence recognition model achieve promising results, outperforming several previous state-of-art methods. Few recent works have posted better accuracy scores than our method, however these methods are more computationally expensive when compared to our model, as can be seen from the parameter numbers. Comparing with [9], our method gives similar performance, however [9] uses OpenPose pose estimation model which increases the model size. Hence, although the work presents trainable parameter count as 62,583, the added overhead of the OpenPose model increases the parameter count to approximately 26 million. Our time analysis, given in Table 7, further supports our claim that our presented model is more lightweight than [9]. The CUE-Net architecture [21] outperforms our model by 4%, however, the parameter count is almost 350 times larger than our model.

Table 1. Comparison of violence recognition accuracy on full length videos of RWF-2000 dataset.

Method	Accuracy (%)	Total Parameters(in millions)
Cheng et al. [6]	87.25	0.272
SPIL [24]	89.3	–
Islam et al. [12]	89.75	0.333
Garcia-Cobo et al. [9]	90.25	26
VD-Net [26]	88.2	4.470
Pratama et al. [20]	90.5	66.6
Ullah et al. [27]	91.15	25
CUE-Net [21]	94.0	354
Ngoc et al. [18]	89.55	4.7
Ours (Teacher EVNet)	89.25	1.19
Ours (Student EEVNet)	90.5	1.19

5.2 Results for Early Violence Prediction

The results of our early violence prediction model on RWF-2000 dataset are shown in Table 2. Further we compare the performance of the models presented in [9,12] on partially observed dataset. It is evident that these models struggle to accurately recognize violent activities from partially observed videos. Especially for observation ratio of 20%, our proposed model records almost 10% and 20% better accuracy scores than [9,12]. The third row shows results of our teacher EVNet model. The fourth row depicts results of our early prediction model (EEVNet) trained with loss presented in [28], which we describe in Eq. 4. Using this approach, we are again able to record better performance for partial videos. However, for later progress levels, the student model loses some of the performance of the teacher model. In this regard, the superior performance of our proposed importance preserving knowledge distillation training can be seen for observation ratios 40%, 60%, 80% and 100%. These higher accuracy scores validate our use of important features from specialized teachers for knowledge distillation. Using important features, we are able to maintain and exceed the performance of the teacher model for later progress levels. We also show early violence prediction results of our model on RLVS dataset in Table 3. For RLVS, we partition our dataset into 80-20% train-test split. This is done because applying a 5-fold cross-validation approach is not suitable for our proposed teacher-student framework. Similar performance improvements can also be seen on RLVS dataset for our proposed methodology.

Table 2. Results of early violence prediction on RWF-2000 dataset.

Method	20%	40%	60%	80%	100%
Islam et al. [12]	70.74	81	82.5	83.25	87
Garcia-Cobo et al. [9]	60	72.5	80	86.25	88.75
EVNet	75.75	83.25	86.5	88.75	89.25
EEVNet with L loss	81.25	84.5	86.5	86.75	87.75
EEVNet with L_{MTI}	83.5	87.25	89.5	90.0	90.5

Table 3. Results of early violence prediction on RLVS dataset.

Method	20%	40%	60%	80%	100%
EVNet	51.5	85.25	90	92	94.25
EEVNet with L loss	85.75	88.75	90.0	91.25	91.75
EEVNet with L_{MTI}	92.25	92.75	92.25	92	92.75

5.3 Ablation Studies

Effect of Squeeze and Excitation Module. The recognition performance of our model with and without our introduced Squeeze and Excitation module is reported in Table 4. Our model using the SESNet achieves a accuracy of 89.25%, whereas the standard ShuffleNetv2 model reports an accuracy of 88%. Further, as seen from the table, our model using SESNet uses only 1,656 more parameters. These results show the effectiveness of our proposed SESNet architecture.

Table 4. Effects of using SESNet.

Method	Accuracy (%)	Total Parameters
EVNet with ShuffleNetv2	88	1197691
EVNet with SESNet	89.25	1199347

Effect of γ on Early Violence Prediction. We list the effects of γ on the early prediction performance using EEVNet in Table 5. The value of γ in Eq. 8 regulates how aggressively the student model is penalized for not aligning it's features with those of the teacher models. With $\gamma = 2$, we are able to achieve the best performance.

Effect of Multi-teacher Importance Preservation. To analyse the effects of our multi-teacher importance preservation, we run the following ablation studies. We train the student model using only classification loss L_C, which is taken as the Baseline model. Next we train the student model with loss L_{MT}, as described

Table 5. Effect of γ on early violence prediction performance.

Value of γ	20%	40%	60%	80%	100%
0.1	82	84.5	87	87.75	88.75
0.5	82.5	86	87.5	87.25	88.75
1.0	81.75	86.5	90	89.5	90.25
2.0	83.5	87.25	89.5	90.0	90.5

in Eq. 5, with $\gamma = 1.0$. Finally we use the L_{MTI} loss described in Eq. 8 with $\gamma = 1.0$. As can be seen from Table 6, our L_{MTI} loss excels in this scenario. This is because the student model learns only the important features from the teacher, allowing non critical features to be modified for other progress levels.

Table 6. Effects of importance preserving knowledge distillation loss.

	20%	40%	60%	80%	100%
Baseline	75	81	84.75	85.5	86.75
Baseline + L_{MT}	82.25	85.25	88.0	88.75	88.5
Baseline + L_{MTI}	81.75	86.5	90	89.5	90.25

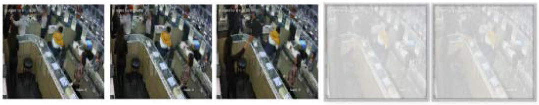
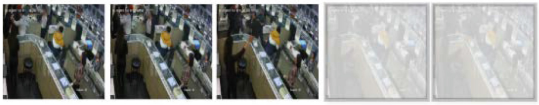
Table 7. Runtime analysis. Average time taken over 50 videos from RWF-2000 dataset.

Method	Total Time (in seconds)
Garcia-Cobo et al.	1.668
Ours	0.378

5.4 Model Efficiency and Time Analysis

For the purpose of real-time recognition, our model should be lightweight and be capable of providing fast predictions. In Table 1, we show the efficiency of our model with the help of parameter count. In this section, we verify the real-time capabilities of our model by measuring the inference time. Following [9], 50 test videos from RWF-2000 dataset are randomly selected. Using a batch size of 1, we process each video independently and then take the average of the inference time across the 50 videos. As shown in Table 7, our model takes 0.378 s to predict violence from full length videos. This demonstrates the capability of our model for providing real-time prediction. We have compared runtime with [9], and our model's inference time is 1.29 s faster. However, this runtime information is not available for other existing work.

Table 8. Qualitative results of our early violence classifier on the RWF-2000 dataset. We show prediction results of teacher and EEVNet on the same video for observation ratios 20%, 40% and 60%. Blurred frames denote unobserved part of the videos.

Video Frames	**Ground Truth**	**Predicted Label**
Teacher EVNet model 	Violence	Non-Violence
Student EEVNet model Observation Ratio = 20%	Violence	Violence
Teacher EVNet model 	Non-Violence	Violence
Student EEVNet model Observation Ratio = 40%	Non-Violence	Non-Violence
Teacher EVNet model 	Violence	Non-Violence
Student EEVNet model  Observation Ratio = 60%	Violence	Violence

5.5 Qualitative Analysis

We show qualitative results of our proposed early violence prediction method on the RWF-2000 dataset in Table 8. Examples from 3 different videos with different observation ratios are shown. For every video, the first row shows the prediction result of our Teacher EVNet model, and the second row presents results of our Student EEVNet on the same observation ratio. We show examples from multiple observation ratios of 20%, 40% and 60%. In the first example, a small amount of initial 30 frames (observation ratio of 20%) are provided to the EVNet and EEVNet models. It is important to note, violence activity has already started at this point. As can be seen, the violence recognition EVNet wrongly classifies the video as Non-Violence. However, for the same partial input from the same video, our early violence prediction network EEVNet accurately predicts the video as violence. The same can be seen for the third example with observation ratio of 60%. Same behaviour is also seen for Non-Violence videos. In the second example, large gathering of people in partial video frames causes the EVNet to misclassify the video as Violence. Overall, for partially observed videos, our EEVNet model correctly predicts the video class, whereas, the EVNet misclassifies.

6 Conclusion

In this paper, we present a novel violence prediction model with the capability to accurately predicting violence from partially observed videos. We design an efficient violence recognition model leveraging our proposed SESNet architecture. The SESNet model is an efficient 3D convolution network enhanced with squeeze and excitation attention modules, making the architecture capable of capturing key spatio-temporal features. Further, we propose a novel multi-teacher importance preserving knowledge distillation approach for training our early violence classifier. Extensive experimental studies on two public datasets show promising results of our method for early violence prediction.

Although given the promising results presented in this work, we are still limited by inadequate dataset for early violence prediction. Most surveillance violence dataset available in the public domain contain short duration videos. Detecting violence a few seconds earlier is not beneficial for real-life deployment. The full potential of such early violence prediction systems can be achieved from long duration videos, where the model predicts violence far in the distant future. Hence, a long duration violence anticipation dataset in surveillance setting is critical for further progress in this field. However, this work acts as an important starting point for further research in the topic.

References

1. Ahn, D., Kim, S., Hong, H., Ko, B.C.: Star-transformer: a spatio-temporal cross attention transformer for human action recognition. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pp. 3330–3339 (2023)
2. Basak, S., Gautam, A.: Diffusion-based normality pre-training for weakly supervised video anomaly detection. Expert Syst. Appl. **251**, 124013 (2024)
3. Bermejo Nievas, E., Deniz Suarez, O., Bueno García, G., Sukthankar, R.: Violence detection in video using computer vision techniques. In: Real, P., Diaz-Pernil, D., Molina-Abril, H., Berciano, A., Kropatsch, W. (eds.) CAIP 2011. LNCS, vol. 6855, pp. 332–339. Springer, Heidelberg (2011). https://doi.org/10.1007/978-3-642-23678-5_39
4. Camporese, G., Coscia, P., Furnari, A., Farinella, G.M., Ballan, L.: Knowledge distillation for action anticipation via label smoothing. In: 2020 25th International Conference on Pattern Recognition (ICPR), pp. 3312–3319. IEEE (2021)
5. Chen, M.Y., Hauptmann, A.: Mosift: recognizing human actions in surveillance videos. Computer Science Department, p. 929 (2009)

6. Cheng, M., Cai, K., Li, M.: Rwf-2000: an open large scale video database for violence detection. In: 2020 25th International Conference on Pattern Recognition (ICPR), pp. 4183–4190. IEEE (2021)
7. Dai, Q., et al.: Fudan-huawei at mediaeval 2015: detecting violent scenes and affective impact in movies with deep learning. In: MediaEval, vol. 1436 (2015)
8. Ding, C., Fan, S., Zhu, M., Feng, W., Jia, B.: Violence detection in video by using 3D convolutional neural networks. In: Bebis, G., et al. (eds.) ISVC 2014. LNCS, vol. 8888, pp. 551–558. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-14364-4_53
9. Garcia-Cobo, G., SanMiguel, J.C.: Human skeletons and change detection for efficient violence detection in surveillance videos. Comput. Vis. Image Underst. **233**, 103739 (2023)
10. Hassanin, M., Anwar, S., Radwan, I., Khan, F.S., Mian, A.: Visual attention methods in deep learning: an in-depth survey. Inf. Fusion **108**, 102417 (2024)
11. Hassner, T., Itcher, Y., Kliper-Gross, O.: Violent flows: real-time detection of violent crowd behavior. In: 2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, pp. 1–6. IEEE (2012)
12. Islam, Z., Rukonuzzaman, M., Ahmed, R., Kabir, M.H., Farazi, M.: Efficient two-stream network for violence detection using separable convolutional lstm. In: 2021 International Joint Conference on Neural Networks (IJCNN), pp. 1–8. IEEE (2021)
13. Kang, M., Park, J., Han, B.: Class-incremental learning by knowledge distillation with adaptive feature consolidation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 16071–16080 (2022)
14. Laptev, I.: On space-time interest points. Int. J. Comput. Vision **64**, 107–123 (2005)
15. Lea, C., Flynn, M.D., Vidal, R., Reiter, A., Hager, G.D.: Temporal convolutional networks for action segmentation and detection. In: proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 156–165 (2017)
16. Li, J., Jiang, X., Sun, T., Xu, K.: Efficient violence detection using 3d convolutional neural networks. In: 2019 16th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS), pp. 1–8. IEEE (2019)
17. Ma, N., Zhang, X., Zheng, H.T., Sun, J.: Shufflenet v2: practical guidelines for efficient cnn architecture design. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 116–131 (2018)
18. Ngoc, H.N., et al.: An efficient approach for real-time abnormal human behavior recognition on surveillance cameras. In: 2023 IEEE 17th International Conference on Automatic Face and Gesture Recognition (FG), pp. 1–6. IEEE (2023)
19. Park, J., Kang, M., Han, B.: Class-incremental learning for action recognition in videos. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 13698–13707 (2021)
20. Pratama, R.A., Yudistira, N., Bachtiar, F.A.: Violence recognition on videos using two-stream 3d cnn with custom spatiotemporal crop. Multimedia Tools Appl. 1–23 (2023)
21. Senadeera, D.C., Yang, X., Kollias, D., Slabaugh, G.: Cue-net: violence detection video analytics with spatial cropping enhanced uniformerv2 and modified efficient additive attention. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 4888–4897 (2024)
22. Simonyan, K., Zisserman, A.: Two-stream convolutional networks for action recognition in videos. In: Advances in Neural Information Processing Systems, vol. 27 (2014)

23. Soliman, M.M., Kamal, M.H., El-Massih Nashed, M.A., Mostafa, Y.M., Chawky, B.S., Khattab, D.: Violence recognition from videos using deep learning techniques. In: 2019 Ninth International Conference on Intelligent Computing and Information Systems (ICICIS), pp. 80–85 (2019)
24. Su, Y., Lin, G., Zhu, J., Wu, Q.: Human interaction learning on 3D skeleton point clouds for video violence recognition. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J.-M. (eds.) ECCV 2020. LNCS, vol. 12349, pp. 74–90. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-58548-8_5
25. Sudhakaran, S., Lanz, O.: Learning to detect violent videos using convolutional long short-term memory. In: 2017 14th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS), pp. 1–6. IEEE (2017)
26. Ullah, F.U.M., et al.: Ai-assisted edge vision for violence detection in iot-based industrial surveillance networks. IEEE Trans. Ind. Inf. **18**(8), 5359–5370 (2021)
27. 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)
28. Wang, X., Hu, J.F., Lai, J.H., Zhang, J., Zheng, W.S.: Progressive teacher-student learning for early action prediction. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 3556–3565 (2019)
29. Zhang, Q.L., Yang, Y.B.: Sa-net: shuffle attention for deep convolutional neural networks. In: ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 2235–2239. IEEE (2021)
30. Zhao, P., Xie, L., Wang, J., Zhang, Y., Tian, Q.: Progressive privileged knowledge distillation for online action detection. Pattern Recogn. **129**, 108741 (2022)

Improving Temporal Action Segmentation and Detection with Hierarchical Task Grammar

Qiu Yihui(✉) and Deepu Rajan

College of Computing and Data Science, Nanyang Technological University,
Singapore, Singapore
qiuy0007@e.ntu.edu.sg
https://www.ntu.edu.sg/

Abstract. Human activities are inherently task-oriented, and integrating explicit task learning into action segmentation models is hypothesized to enhance performance. However, empirical evaluations using the Ego4D Goal Step dataset reveal a paradox: the inclusion of learning tasks deteriorated model performance. This issue partially arises from limited task samples, i.e., over 50% of tasks have less than two training samples, leading to bias and overfitting in the training process. To address this, we propose a novel grammar induction method to accurately capture the hierarchical decomposition of a task with limited task samples, and use the induced grammar to guide neural predictions. Experiments demonstrate that our induction method achieves comparable results with SOTA on the Breakfast dataset with as few as two training samples for each task. Additionally, incorporating our grammar significantly boosts temporal action segmentation results for Ego4D Goal Step dataset by 8%. This approach not only mitigates data scarcity but also enhances the robustness and accuracy of action segmentation and action detection models.

Keywords: Temporal Action Segmentation · Detection · Grammar Model

1 Introduction

Human activities are inherently goal-oriented, and are influenced by the subject's goals, object manipulation and interaction in the environment. Learning the compositional structure of human activity poses a significant challenge in video understanding research. Temporal action segmentation and action detection cover a critical aspect of this domain, aiming to temporally segment an untrimmed video and label each segment with predefined action labels [7,8].

The primary distinction between the two tasks lies in their objectives. Temporal action segmentation involves labeling each video frame with an action class. In contrast, action detection is concerned with identifying and localizing specific

A. Antonacopoulos et al. (Eds.): ICPR 2024, LNCS 15329, pp. 196–211, 2025.
https://doi.org/10.1007/978-3-031-78110-0_13

instances of actions in an untrimmed video. Consequently, temporal action segmentation is typically evaluated on videos that are instructional in nature, e.g., Breakfast [32], and use metrics such as edit distance, frame-wise accuracy, and Intersection over Union (IoU). Conversely, temporal action detection is commonly evaluated on non-instructional videos such as Ego4D [33] and use metrics like average mean Average Precision (average mAP).

Motivated by the success of deep neural networks in understanding and generating long-range text sequences, recent methods for temporal action segmentation and detection implicitly learn temporal relations of actions [8,34]. However, they struggle with long-term human action sequences. A popular hypothesis is that learning goal/task of the action sequence can help the model capture long-term action dependencies better. However, empirical evaluations using the Ego4D Goal Step dataset [1], one of the largest human activity video datasets available, reveal that learning task labels with action labels can degrade performance in temporal action segmentation tasks, as we show in Sect. 4.3. This issue arises partially from the limited number of task samples; over 50% of tasks have fewer than two training samples, leading to bias and overfitting.

Two main biases emerge from limited task samples: ordering bias and missing actions bias. Ordering bias occurs when the probability of action A following action B is significantly higher than the reverse, even though both orders are equally likely in reality. For example, in a "make coffee" task, the training data might show "pour coffee" before "pour milk," although these actions can occur in any order. Missing action bias arises because a video may start or end at any stage of the task, thus not containing the full sequence of actions required to complete it. For instance, a "take milk from fridge" action might be absent in a cereal-making video if the milk is already on the table.

To address these limitations, we propose a novel Hierarchical Task Grammar induction algorithm. A grammar captures rules to represent hierarchical temporal structure of a sentence in formal language. In our case, the rules describe how to decompose and order actions that achieve a given task. Our method focuses on object-centered action transitions, decomposing tasks into object interactions. The induced grammar mitigates ordering biases and incorporates missing actions using universal object graphs, enhancing generalization capabilities. Additionally, we introduce a graph search-based algorithm for task inference and confidence score refinement that accommodates out-of-grammar action prediction.

The main contributions of this paper can be summarized as follows:

- We introduce a novel grammar induction algorithm that significantly reduces biases introduced by limited task samples.
- We develop a graph search based algorithm for task inference and confidence score refinement for temporal action segmentation and detection, that works for both instructional and non-instructional datasets.
- We show that the proposed method significantly improves the performance of temporal action segmentation and detection models, as demonstrated through

a comprehensive evaluation on four benchmarks datasets: 50 Salads [35], Breakfast [32], Ego4D Goal-Step [1] and Epic-Kitchens-100 [36].

2 Related Work

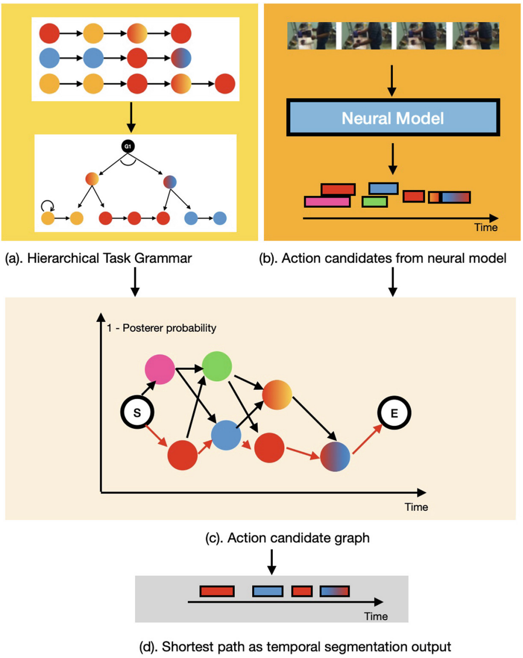
Temporal action segmentation and detection require models that effectively capture the sequential relationships of actions. The mainstream approach involves deep learning-based temporal models, such as RNNs [17,18], TCNs [12–14], and Transformers [15,16], which facilitate information exchange across frame-wise features and have achieved notable results. Techniques that enhance context learning, such as expanding the temporal receptive field [28–30], aggregating features over multiple granularity [26], and adapting attention mechanisms [27], have shown further improvements. Deep learning based methods, however, still struggle with forgetting issues and heavy computation burden to capture the full context of a video [11].

Human action sequences are naturally task-oriented and hence follow certain ordering constraints. Many methods have been proposed to integrate additional modules that encapsulate high level semantics to guide predictions from neural models. Kuehne et al. [19] proposed combining a framewise RNN model with a coarse probabilistic inference where action sequence are modeled by hidden Markov models (HMMs). Huang et al. [9] modeled relation of multiple action segments in various time spans by using Graph Convolution Networks. Ahn and Lee [10] proposed a refinement model that implicitly learn temporal action relation from hierarchical video representations. Xu et al. [11] proposed Differentiable Temporal Logic (DTL), that introduces temporal constraints to deep networks.

Our work aligns with **grammar parsing**, which represent hierachical task structure using grammar induced from sample task sequences. Grammar is first used for action anticipation in [24,25]. The AND-OR grammar was learnt from example strings of symbols, each representing an action according to the grammar's language. However, the method can only processes deterministic inputs. Vo and Bovick [20] proposed to use stochastic context-free grammar which takes in probabilistic sequence inputs instead of deterministic symbolic inputs. Qi et al. [3] utilized the ADIOS grammar induction algorithm to induce grammar. Recently Gong et al. [2] proposed a grammar induction algorithm based on key actions and temporal dependencies, considering recursive temporal structures. However, existing grammars are either hand-crafted, which is costly, or they rely on the action sequences in the training data, thereby inheriting the biases caused by limited task samples. Our proposed grammar induction method is able to minimize such biases with a small number of task samples.

3 Method

Our approach can be considered as a neuro-symbolic one, where the Hierarchical Task Grammar (HTG) guides neural predictions. Given a detected action candidate from neural model, the posterior probability of the action given the task and observation is calculated considering both the task prior from HTG and detection confidence from neural models.



(a). Hierarchical Task Grammar

(b). Action candidates from neural model

(c). Action candidate graph

(d). Shortest path as temporal segmentation output

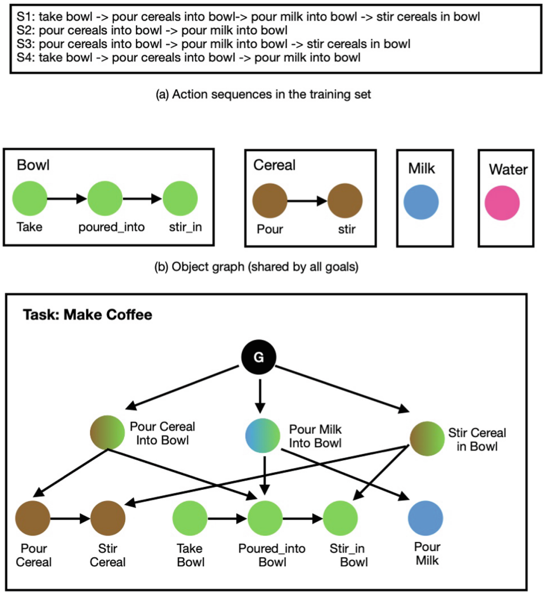
Fig. 1. Overall pipeline of the proposed method: (a) HTGs are constructed from action sequences for each task from the training set. (b) Construct a graph based on action candidates from off-the-shelf temporal action segmentation model (c) Update edge weights with confidence score from neural model and task prior from Hierarchical Task Grammar, and find the shortest path for each task. (d) the nodes of shortest path are use as output of action segmentation.

The overall pipeline of the proposed method consists of three steps, as illustrated in Fig. 1. Firstly, HTGs are constructed from action sequences for each task from the training set. Secondly, off-the-shelf temporal action segmentation/detection model outputs action candidates in the format {action, start_time, end_time, confidence}. Thirdly, a directed acyclic graph is constructed based on the action candidates, where the edge weight is calculated by combining confidence score from the neural model and the task prior from task grammar of a task G_e. The task inference is then conceptualized as a shortest path problem, aiming to identify a path from the start to the end node that minimizes the average weights of the edges involved. The average weights of the shortest path is assigned as the task score G_{score} for the task G_e. Lastly the weights for the task with minimum task score is the output for temporal action detection. The optimal path, associated with the task, constitutes the refined output for temporal action segmentation.

3.1 Hierarchical Task Grammar (HTG) Induction

We aim to develop a task grammar induction algorithm that accurately represents and decomposes a given task, capturing the actions and their ordering, even with small number of sample action sequence. Two primary challenges are task decomposition and action ordering.

The first challenge is **task decomposition**. A straightforward approach to construct task grammar is to break down a task into sub-tasks, and then decompose them further until the atomic action level. However, a drawback of this method is the lack of a standardized way for defining and decomposing sub-tasks, and usually requires the sub-task to be manually crafted for each task. For instance, task "make coffee" can be decomposed into sub-task "prepare machine and tools", "prepare coffee powder", "prepare hot water", "mix coffee and water" or more generally "prepare ingredients", "mix ingredients". To address this issue, we propose using action interactions as sub-tasks, which is easy to construct, and applicable across various tasks and datasets.



S1: take bowl -> pour cereals into bowl-> pour milk into bowl -> stir cereals in bowl
S2: pour cereals into bowl -> pour milk into bowl
S3: pour cereals into bowl -> pour milk into bowl -> stir cereals in bowl
S4: take bowl -> pour cereals into bowl -> pour milk into bowl

(a) Action sequences in the training set

(b) Object graph (shared by all goals)

(c) Task graph

Fig. 2. Example of HTG construction. (a) Example action sequences for task "make cereal". Assume S1 is the complete action sequence. (b) object-centered graph shared by all the tasks (c) Task graph constructed.

The second challenge is **action ordering**. Learning the correct order of actions is challenging, primarily due to biases introduced by limited training action sequences. We observed that action transitions within a single object exhibit greater consistency across tasks. For instance, the actions on a fridge typically follow the order "open" → "take/put something" → "close", which is consistent across tasks and even datasets. With this observation, we propose constructing object graphs for each object, which serve as foundational components applicable across all tasks. By aggregating object graphs into task graph, with a small number of sample sequences, we can recover missing actions to a great extent. We will explain and demonstrate with an example in Fig. 3 at the end of this section.

Next we will first explain the components and construction of HTG, with an example shown in Fig. 2. Assume for the make cereal task, there are four task samples as shown Fig. 2(a).

Object Graph is constructed for each object by extracting only the actions and their transition related to that object from the task samples. It contains both atomic actions and interactions involving the object as nodes, as shown in Fig. 2(b). For interactions involving multiple objects, such as 'pour cereal into bowl', the action for 'cereal' is denoted as 'pour', and for 'bowl' as 'poured into' within their respective object graphs. This approach enhances the robustness and generality of transition probabilities by consolidating similar actions. For example 'bowl' has a single 'poured into' node representing all actions like 'pour water into bowl' and 'pour milk into bowl'.

Directed edges then connect these action nodes if one action typically follows another. We use transition and inverse conditional probabilities to capture the conditional probability between two connected nodes. For a connected edge from node a_k to a_j ($a_k \rightarrow a_j$) in object graph for object o, the transition probability $P_o(a_j|a_k)$ quantifies the likelihood of transitioning from action a_k to action a_j and is computed as

$$P_o(a_j|a_k) = \frac{\text{Number of transitions from } a_k \text{ to } a_j}{\text{Total number of transitions from } a_k}. \tag{1}$$

The posterior conditional probability $P_o(a_k|a_j)$ represents the probability of action a_k given that the current action is a_j and can be computed through Bayes theorem.

Task Graph is constructed from instances of action sequences for the task. Figure 3 demonstrates construction of task graph step by step. Starting from a root node G (the highest level), we examine a given action sequence ["pour cereal into bowl" → "pour milk into bowl"]. Upon identifying object interaction, we connect them to the G node as shown in Fig. 3(a). In this example both actions "pour cereal into bowl" and "pour milk into bowl" are interactions. We then integrate object graphs of objects involved in the interactions into the task graph, connecting relevant nodes. For example in Fig. 3(b), linking the "pour cereal into bowl" interaction to the "pour" node in the cereal object graph and the "poured into" node in the bowl object graph). This process continues for subsequent interactions in the sequence, ensuring they are either integrated or expanded as necessary until all interactions are included in the task graph. Next, we examine the atomic actions within the action sequence and verify their presence in the task graph. If absent, we add the respective object graph and connect the atomic action to the root node. Lastly, for all nodes connected to the root node, the weight of the edge is the conditional probability $P(a^I|G_e)$, that represents the likelihood of a interaction a^I occurring given the task G_e.

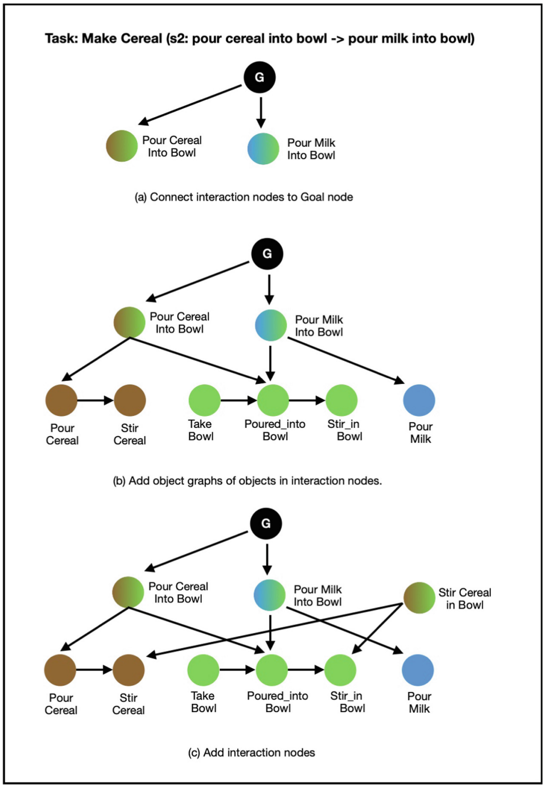


Fig. 3. Example of HTG construction. Task graph constructed with s2. Even there are only two actions in the sequence, our method is able to capture "take bowl" and "stir cereals in bowl" from shared object graph.

Next we compare this task graph with one using a single action sequence to construct the task graph for "make cereal". Assume s1 is the complete sequence for the task that contains four actions "take bowl" → "pour cereal into bowl" → "pour milk into bowl" and "stir cereals in bowl". Figure 2(c) shows the task graph generated with s1. Figure 3(c) shows the task graph generated with s2, which only contains two actions "pour cereal into bowl" and "pour milk into bowl". With only two actions in the sample sequence, our method is able to capture the missing actions “take bowl” and “stir cereals in bowl” from object graphs.

3.2 Action Candidates Graph

The action candidates graph is a directed acyclic graph that is constructed with action candidates from an off-the-shelf temporal action segmentation/detection model, as shown in Fig. 1(c). Each action candidate, which is defined by {action, start_time, end_time, confidence}, serves as a node. Directed edges are established between two nodes if the start time of one node occurs within a specified time margin (+ve/–ve) from the end time of another. Positive margin allows gap between two actions and negative margin allows overlap between two actions.

3.3 Action Posterior Probability Formulation

Grammar Prior. We have introduced three probabilities in HTG: the conditioned probability of an interaction given a task $P(a^I|G_e)$ and the transition probability $P_o(a_j|a_k)$ and inverse condition probability $P_o(a_k|a_j)$ of connected actions for object o in object graph.

The task prior $p(a \mid G_e)$ is decomposed to $p(a \mid a^I)$, which is the probability of action a given interaction a^I, and $p(a^I \mid G_e)$, which is the probability of interaction a^I given task G_e. Thus,

$$p(a \mid G_e) = p_o(a \mid a^I) \cdot p(a^I \mid G_e) \tag{2}$$

If a is an action that leads to an interaction a^I, $p(a \mid a_i^I)$ is calculated using chain rule on transition probability, where a_0 to a_N represents the path from node a to a^I ($a_0 = a$, $a_N = a^I$). Otherwise, inverse condition probability is used, where a_0 to a_N represents the path from node a^I to a ($a_0 = a^I$, $a_N = a$),

$$p_o(a \mid a^I) = \prod_{i=0}^{N} p_o(a_i \mid a_{i-1}) \tag{3}$$

Action Likelihood. $p(\Gamma_a \mid a)$, which represents the likelihood of action a given observation Γ_a, is the confidence score of an action candidate from the neural model.

Posterior Probability. The posterior probability of an action candidate given the observation Γ_a and task G_e is formulated as product of the likelihood $p(\Gamma_a \mid a)$ and the task prior probability $p(a \mid G_e)$. Following Bayes theorem,

$$\begin{aligned} p(a \mid \Gamma_a, G_e) &\propto p(\Gamma_a \mid a) \cdot p(a \mid G_e) \\ &= p(\Gamma_a \mid a) \cdot p_o(a \mid a^I) \cdot p(a^I \mid G_e) \\ &= p(\Gamma_a \mid a) \cdot \prod_{i=0}^{N} p_o(a_i \mid a_{i-1}) \cdot p(a^I \mid G_e) \end{aligned} \tag{4}$$

3.4 Time-Normalized Dijkstra Algorithm

Traditional grammar parsing algorithms are grammar-centered, restricting outputs to actions defined within the grammar structure. This limitation means that actions not explicitly tied to a task will never be selected, even if they receive high confidence scores from a neural model. This causes issue when applied to non-instructional videos as irrelevant actions appear intermediately. To address this issue, we reformulate the action sequence generation as a shortest path problem on the action candidates graph. In cases where multiple nodes exist for a particular time frame, the algorithm selects the node with the lowest weight, irrespective of whether the action conforms to the grammar.

As we are going to use the shortest path algorithm on the action candidates graph, an edge pointing from node i to node j is assigned a **weight** $w_{ij} = 1 - p(a_j|\Gamma_{a_j}, G_e)$ where $p(a_j|\Gamma_{a_j}, G_e)$ is the action posterior probability of the node that the edge points to. If an action candidate is not defined in the task graph, we assign a dummy $p(a \mid G_e) = 1 \times 10^{-3}$ for it.

Dijkstra's Algorithm is a well-known method for finding the shortest path between nodes in a graph with non-negative edge weights. In our action candidate graph, each edge has a weight value between 0 to 1, and the number of nodes in a path from start node to end node may vary. i.e. a path connecting nodes representing longer duration will have less nodes. The original Dijkstra's algorithm finds the shortest path with summed weights, i.e. less nodes leads to less total weight. Hence this algorithm tends to favor paths with fewer edges (nodes) over paths with more edges, which can lead to sub-optimal paths.

To address this issue, we propose a modified version of Dijkstra's algorithm, called **Time-normalized Dijkstra Algorithm**, that considers the average weight of edges in the path instead of the total weight. This modification ensures that paths with short actions are not unfairly penalized.

3.5 Task Inference and Confidence Score Refinement

Edge weights are recalculated for each task $g \in G$ based on the task's Hierarchical Task Grammar (HTG). The task score is the average weight of the shortest path, determined by the Time-normalized Dijkstra Algorithm. The weights of the task with minimum task score are the refined confidence scores for action detection. The shortest path of the task provides the refined output for temporal action segmentation.

4 Experiments

4.1 Datasets and Evaluation Metrics

Datasets. For temporal action segmentation, we conduct experiments on two instructional datasets: **50Salads** and **Breakfast**. The **50Salads** dataset consists of 50 egocentric videos depicting individuals preparing salads, featuring 17 fine-grained actions performed by 25 participants. The **Breakfast** dataset includes

1,712 videos capturing 52 individuals preparing 10 different Breakfast activities across 18 kitchens, with 48 actions performed.

For temporal action detection, we conduct experiments on the challenging non-instructional datasets **Ego4D Goal Step dataset** and **EPIC-Kitchens-100**. The **Ego4D Goal Step dataset** [1] is a subset of Ego4D dataset, comprising over 3,670 h of egocentric video footage, annotated with dense procedural step segments totaling 48,000 annotations and high-level task annotations spanning 2,807 h. Note that as the test dataset is not released, we randomly split the validation set into new validation set and test set with one to one ratio, and report the result over 8 runs. The **EPIC-Kitchens-100** dataset features 100 videos of daily activities in kitchens, including 300 annotated objects and 3,805 actions. Note that EPIC-Kitchens-100 does not contain task annotations. We challenge our method's adaptability to apply Hierarchical Task Grammar generated from Ego4D directly to Epic-Kitchen-100.

Evaluation Metrics . For temporal action detection, we report mAP averaged over action categories, and over temporal IoUs 0.1, 0.2, 0.3, 0.4, 0.5. For temporal action segmentation, we report edit score, F1@10, 25, 50 scores, and frame-wise accuracy.

4.2 Neural Models

For temporal action segmentation task, we use ASFormer [15] based on Transformer and MS-TCN [14] based on CNNs, following previous work [2]. For temporal action detection task, we use ActionFormer [16] and EgoOnly [31] as our neural model due to its wide adoption and the availability of open-source implementations following [1].

4.3 Results

Temporal Action Segmentation. Table 1 and Table 2 show the performance of applying the proposed method to temporal action segmentation task for 50Salads and Breakfast. The comparison between the neural model and after refinement reveals significant improvements in both edit scores and F1 scores. Our method outperforms the state-of-the-art grammar parsing methods across all metrics.

Temporal Action Detection. Table 3 shows the performance of applying the proposed method to the action detection task for the Ego4D Goal Step dataset. We use the essential steps annotated in the task samples to construct HTGs. The first row presents results from ActionFormer trained solely on action labels, while the second row includes training on both action and task labels. As previously noted, over 50% of tasks in the Ego4D Goal Step dataset have fewer than two samples in the training set. The results indicate that directly training on task labels with such limited samples deteriorates the model's performance.

Table 1. The temporal action segmentation performance comparison on 50Salads.

Model	Refinement algo	Edit	F1			acc.
			@10	@25	@50	
ASFormer [15] (reproduced)	–	76.5	83.8	81.7	74.8	**86.1**
	ADIOS-OR [3]	61.1	72.0	70.1	62.4	78.9
	KARI [2]	79.9	85.4	83.8	77.4	85.3
	ours	**83.9**	**87.4**	**86.6**	**78.2**	85.4
MS-TCN(reproduced)	–	62.4	69.5	65.3	55.7	75.2
	ADIOS-OR [3]	61.9	69.1	66.9	57.2	74.2
	KARI [2]	66.7	75.1	73.2	60.8	76.7
	ours	**69.2**	**78.8**	**75.9**	**64.2**	78.3

Table 2. The temporal action segmentation performance comparison on Breakfast

Model	Refinement Algorithm	edit	F1			acc.
			@10	@25	@50	
ASFormer [15] (reproduced)	–	75.6	77.3	70.2	59.4	**74.3**
	ADIOS-OR [3]	70.3	71.8	66.8	54.2	71.8
	KARI [2]	77.8	78.8	73.7	60.8	74.0
	ours	**78.1**	**79.2**	**74.0**	**61.1**	74.1
MS-TCN (reproduced)	–	69.7	70.7	65.1	52.6	**69.4**
	ADIOS-OR [3]	69.6	69.6	64.3	50.3	68.2
	KARI [2]	74.9	74.6	68.7	55.1	68.8
	ours	**75.2**	**74.8**	**68.9**	**55.2**	69.3

Conversely, our model better captures the task step structure, thereby enhancing the model's performance.

The EPIC-Kitchens-100 dataset lacks task annotations. To challenge the generalizability of our method, we applied HTGs constructed from the Ego4D Goal Step dataset directly to the EPIC-Kitchens-100 dataset. The results, shown in Table 4, demonstrate that our method consistently boosts the model's performance, albeit by a small margin.

4.4 Ablation Study

We conducted three ablation studies to evaluate our method's key components.

Firstly, experiments on the Breakfast dataset aimed to assess our model's capacity in constructing HTGs using a limited number of task samples. For each task, n samples were randomly picked for each task from training set to construct HTGs, followed by evaluation on the validation set across 8 rounds of random selection. Our findings, detailed in Table 5, demonstrate that our method achieves performance comparable to state-of-the-art grammar parsing algorithms even with as few as two sample action sequences.

Secondly, an ablation study on non-instructional dataset examined the impact of using essential versus all actions to construct HTGs. Constructing accurate grammar from non-instructional video is very challenging as there are considerable amount of optional and irrelevant actions in a task sample. Ego4D Goal Step dataset labels each step of a task sample into either essential, optional or irrelevant. We experiment constructing HTGs based on three setups: essential, essential and optional actions, and all actions. The results in Table 6 shows that using essential actions to construct HTGs is most effective with non-instructional videos.

Table 3. The temporal action detection performance comparison on Ego4D Goal Step dataset.

Model	Average mAP	
	Val	Test
ActionFormer (Action)	9.7	8.5
ActionFormer (Task + Action)	9.5	8.3
ActionFormer (Action) + HTGs (ours)	**10.3**	**9.2**
EgoOnly (Action)	13.0	13.8
EgoOnly (Task + Action)	12.7	13.5
EgoOnly (Action) + HTGs (ours)	**13.8**	**14.3**

Table 4. The temporal action detection performance comparison on Epic-Kitchens-100 dataset.

Model	Verb mAP	Noun mAP
ActionFormer	23.5	21.9
ActionFormer + HTGs (ours)	**23.6**	**22.1**
EgoOnly	21.36	20.95
EgoOnly + HTGs (ours)	**21.6**	**21.1**

Lastly, we compared the performance on the Ego4D Goal Step dataset using the original Dijkstra algorithm versus our proposed time-normalized Dijkstra algorithm. Table 7 illustrates that while the original Dijkstra algorithm led to a deterioration in model performance, our time-normalized variant significantly enhanced performance.

4.5 Qualitative Analysis

Figure 4 presents a visual representation of the refined segmentation results on Breakfast datasets. HTG allows a more flexible temporal structure between actions by removing incorrect ordering of actions due to biases. For instance, in training dataset, action “putting salt and pepper” always happen before “putting egg into pan”. HTG successfully capture the possibility of “putting salt and pepper” happening after “putting egg into pan”. Moreover, similar to other grammar based refine method, our method correctly prioritizes task-related actions, such as putting the egg on the plate over putting the pancake on the plate.

Table 5. Ablation: model performance with different number of available training samples

Model	Refinement Algo	Number of Samples per Task	Edit	F1			Acc.
				@10	@25	@50	
ASFormer [15]	–	All training samples (25 per task)	75.6	77.3	70.2	59.4	**74.3**
	KARI [2] (SOTA)	All training samples (25 per task)	77.8	78.8	73.7	60.8	74.0
	HTGs (ours)	all training samples (25 per task)	**78.1**	**79.2**	**74.0**	**61.1**	74.1
	HTGs (ours)	2 training samples per task	77.7	78.9	73.6	60.9	74.1
	HTGs (ours)	1 training samples per task	76.7	78.0	72	60.1	74.0

Table 6. Ablation: Comparison of performance with different HTG sources

HTG source	Average mAP	
	Val	Test
Essential Steps	**10.3**	**9.2**
Essential + steps	9.5	8.2
All steps	9.4	8.1

Table 7. Ablation: Comparison of performance with original and Time-normalised Dijkstra Algorithm

Dijkstra Algorithm (DA)	average mAP	
	Val	Test
Time-normalised DA	**10.5**	**9.4**
Original DA	8.5	6.4

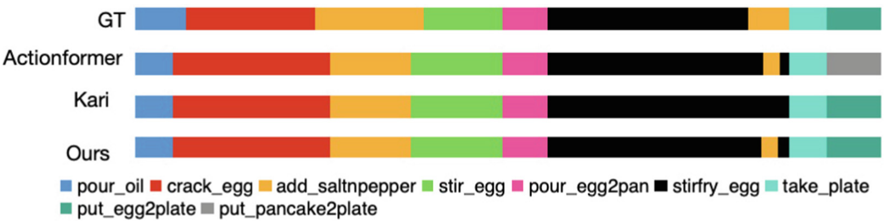


Fig. 4. Qualitative results. HTG correctly preserves the rare action order by putting salt and pepper after putting the egg in the pan, and prioritizes task-related actions, such as putting the egg on the plate over putting the pancake on the plate.

5 Conclusion

We have demonstrated that the proposed approach enhances sequence prediction and uncovers its compositional structure, significantly improving temporal action segmentation and detection in both performance and interpretability. However, our method assumes each video contains only a single task, which may not always be the case. Disentangling multiple tasks remains an intriguing direction for future research.

References

1. Song, Y., Byrne, E., Nagarajan, T., Wang, H., Martin, M., Torresani, L.: Ego4d task-step: toward hierarchical understanding of procedural activities. In: Advances in Neural Information Processing Systems, vol. 36 (2024)
2. Gong, D., Lee, J., Jung, D., Kwak, S., Cho, M.: Activity grammars for temporal action segmentation. In: Advances in Neural Information Processing Systems, vol. 36 (2024)
3. Qi, S., Jia, B., Huang, S., Wei, P., Zhu, S.C.: A generalized earley parser for human activity parsing and prediction. IEEE Trans. Pattern Anal. Mach. Intell. **43**(8), 2538–2554 (2020)
4. Qi, S., Huang, S., Wei, P., Zhu, S.C.: Predicting human activities using stochastic grammar. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 1164–1172 (2017)
5. Chomsky, N.: A note on phrase structure grammars. Inf. Control **2**(4), 393–395 (1959)
6. Backus, J.: Can programming be liberated from the von Neumann style? a functional style and its algebra of programs. Commun. ACM **21**(8), 613–641 (1978)
7. Richard, A., Kuehne, H., Gall, J.: Weakly supervised action learning with rnn based fine-to-coarse modeling. In: Proceedings of the IEEE CVPR, pp. 754–763 (2017)
8. Hu, K., Shen, C., Wang, T., et al.: Overview of temporal action detection based on deep learning. Artif. Intell. Rev. **57**, 26 (2024)
9. Huang, Y., Sugano, Y., Sato, Y.: Improving action segmentation via graph-based temporal reasoning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 14024–14034 (2020)
10. Ahn, H., Lee, D.: Refining action segmentation with hierarchical video representations. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 16302–16310 (2021)
11. Xu, Z., Rawat, Y., Wong, Y., Kankanhalli, M.S., Shah, M.: Don't pour cereal into coffee: differentiable temporal logic for temporal action segmentation. In: Advances Neural on Information Processing Systems, vol. 35, pp. 14890–14903 (2022)
12. Lei, P., Todorovic, S.: Temporal deformable residual networks for action segmentation in videos. In: CVPR, pp. 6742–6751 (2018)
13. Li, S., Farha, Y.A., 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)
14. Farha, Y.A., Gall, J.: MS-TCN: multi-stage temporal convolutional network for action segmentation. In: CVPR, pp. 3575–3584 (2019)

15. Yi, F., Wen, H., Jiang, T.: ASFormer: transformer for action segmentation. In BMVC (2021)
16. Zhang, C.L., Wu, J., Li, Y.: Actionformer: localizing moments of actions with transformers. In: ECCV 2022, pp. 492–510. Springer, Cham (2022)
17. Lin, T., Zhao, X., Fan, Z.: Temporal action localization with two-stream segment-based RNN. In: 2017 IEEE International Conference on Image Processing (ICIP), pp. 3400–3404. IEEE (2017)
18. Richard, A., Kuehne, H., Gall, J.: Weakly supervised action learning with rnn based fine-to-coarse modeling. In: Proceedings of the IEEE CVPR 2017, pp. 754–763 (2017)
19. Kuehne, H., Richard, A., Gall, J.: A hybrid rnn-hmm approach for weakly supervised temporal action segmentation. IEEE Trans. Pattern Anal. Mach. Intell. **42**(4), 765–79 (2018)
20. Vo, N.N., Bobick, A.F.: From stochastic grammar to bayes network: probabilistic parsing of complex activity. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2641–2648 (2014)
21. Richard, A., Kuehne, H., Gall, J.: Action sets: weakly supervised action segmentation without ordering constraints. In: Proceedings of the IEEE CVPR, pp. 5987–5996 (2018)
22. Qi, S., Huang, S., Wei, P., Zhu, S.-C.: Predicting human activities using stochastic grammar. In: Proceedings of IEEE International Conference on Computer Vision, pp. 1173–1181 (2017)
23. Qi, S., Jia, B., Zhu, S.-C.: Generalized earley parser: bridging symbolic grammars and sequence data for future prediction. In: Proceedings of IEEE International Conference on Machine Learning, pp. 4168–4176 (2018)
24. Pei, M., Jia, Y., Zhu, S.C.: Parsing video events with goal inference and intent prediction. In: ICCV, pp. 487–494. IEEE (2011)
25. Si, Z., Pei, M., Yao, B., Zhu, S.-C.: Unsupervised learning of event and-or grammar and semantics from video. In: ICCV (2011)
26. Sener, F., Singhania, D., Yao, A.: Temporal aggregate representations for long-range video understanding. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J.-M. (eds.) ECCV 2020. LNCS, vol. 12361, pp. 154–171. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-58517-4_10
27. Tang, Y., Zhang, X., Ma, L., Wang, J., Chen, S., Jiang, Y.G.: Non-local netvlad encoding for video classification. In: Proceedings of ECCV (2018)
28. Ding, L., Xu, C.: Weakly-supervised action segmentation with iterative soft boundary assignment. In: Proceedings of the IEEE CVPR (2018)
29. Lei, P., Todorovic, S.: Temporal deformable residual networks for action segmentation in videos. In: Proceedings of the IEEE CVPR (2018)
30. Singhania, D., Rahaman, R., Yao, A.: C2F-TCN: a framework for semi-and fully-supervised temporal action segmentation. IEEE Trans. Pattern Anal. Mach. Intell. **45**(10), 11484–11501 (2023)
31. Wang, H., Singh, M.K., Torresani, L.: Ego-only: egocentric action detection without exocentric transferring. In Proceedings of the IEEE CVPR (2023)
32. 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 CVPR (2014)
33. Gramman, K., et al.: Ego4d: around the world in 3,000 hours of egocentric video. In: Proceedings of the IEEE CVPR (2022)
34. Ding, G., Sener, F., Yao, A.: Temporal action segmentation: an analysis of modern techniques. IEEE Trans. Pattern Anal. Mach. Intell. (2023)

35. Stein, S., McKenna, S.J.: Combining embedded accelerometers with computer vision for recognizing food preparation activities. In: Proceedings of the 2013 ACM International Joint Conference on Pervasive and Ubiquitous Computing, pp. 729–738 (2013)
36. Damen, D., et al.: Rescaling egocentric vision: collection, pipeline and challenges for epic-kitchens-100. Int. J. Comput. Vis. 1–23 (2022)

Hybrid Human Action Anomaly Detection Based on Lightweight GNNs and Machine Learning

Miao Feng and Jean Meunier(✉)

Department of computer science and operations research, University of Montreal,
Montreal, Canada
{miao.feng,jean.meunier}@umontreal.ca

Abstract. Detecting human abnormal actions, such as falling, can reduce the cost for both economic and health systems, as well as protect the individuals involved. However, training an anomaly detector from scratch can be computationally expensive. We propose reusing the representations learned by lightweight Graph Neural Networks (GNNs) from a multi-action classification task for quick and accurate anomaly detection. Specifically, the representations are extracted by a lightweight GNNs obtained through self-distillation and are then used for unsupervised anomaly detection with effective machine learning methods such as Gaussian Mixture Model, Dirichlet Process Mixture Model, Isolation Forest, and Local Outlier Factor. This hybrid approach is evaluated on the NTU RGB+D and a subset of Kinetics400 datasets across 20 anomaly detection tasks, including speed-sensitive tasks, exercise actions, and daily activities. The feature extractors used are ST-GCN and AAGCN. Our method achieves an average AUC improvement of 11.4% compared to the unsupervised top model GEPC across the anomaly detection tasks on NTU RGB+D dataset. It also outperforms another supervised method on the URFall dataset. The representations obtained through self-distillation were superior in 8 out of 10 anomaly detection tasks on NTU RGB+D. Additionally, the lightest AAGCN model, which is around 60% lighter than the heaviest model, shows similar or superior performance on average across all anomaly detection tasks for both datasets. However, the representations extracted are usually redundant for anomaly detection, providing more information than anomaly detection needs because their extractors were trained on multi-action classification tasks. Consequently, we also did experiments across all tasks to show that feature reduction improves even more detection performance by up to 17.69% on the subset of Kinetics400 and up to 24.45% on NTU RGB+D.

Keywords: Human Action Recognition · Anomaly Detection · Lightweight Graph Neural Networks · Self-distillation

The authors would like to thank the China Scholarship Council (CSC) and the Natural Sciences and Engineering Research Council of Canada (NSERC) for their financial support.

A. Antonacopoulos et al. (Eds.): ICPR 2024, LNCS 15329, pp. 212–228, 2025.
https://doi.org/10.1007/978-3-031-78110-0_14

1 Introduction

Human action anomaly detection automatically identifies abnormal human actions, which is particularly meaningful for health systems, education, security work, and other areas. For instance, speed-sensitive task detection, such as fall detection, could offer immediate medical service to elderly individuals who live alone. This prompt response can save lives, considering the large number of elderly people who experience falls at least once every year [19]. Additionally, detecting exercise activities such as cycling and skiing can reveal when an exerciser stops and ends up in a dangerous situation, while detecting actions like kicking and fighting can prevent school bullying and improve public safety. The applications of human action anomaly detection are vast, and a quick and relatively accurate anomaly action detector can efficiently prevent harm to individuals and society, thereby reducing associated costs.

Human action recognition (HAR) methods based on deep learning (DL) can extract human action representations (features) into lower dimensions rather than keeping the original large dimensions of the inputs, which is a sequence of action observations. If the extracted representations accurately reflect semantic differences as geometric distances, then the abnormal actions will have representations that are significantly different from normal actions, allowing for their recognition. Recently, graph neural networks (GNNs) have become prominent for extracting human action representations because they specialize in discovering the intrinsic relationships between human joints. These methods represent human skeletons as graphs, with each joint as a graph node and each bone as a graph link. The graph format not only allows for the exploration of human joint topology but also protects personal privacy by removing facial identities and home environments and reduces the dimensionality of the input, unlike RGB videos. The earliest milestone in this field is ST-GCN [25] (Fig. 1). Since then, multiple works based on ST-GCN, such as AAGCN [21], have emerged.

In real-life, a popular anomaly detector is preferred to be light enough to ensure quick anomaly detection without significantly compromising accuracy. Besides, the anomaly detector is expected to distinguish unseen anomalies even if they were not provided to the detector during training. In practice, researchers usually propose end-to-end heavyweight detectors, arguing that their high model capacity guarantees superior detection performance. However, these models are computationally expensive and memory intensive. Moreover, the definition of anomalies can vary significantly across different contexts, even within the same dataset. For instance, even if the extractor is trained from scratch on kicking or falling task and have already captured the semantic distances of their corresponding actions, they may fail on fighting task because they do not take advantage from fighting actions during training. Taking this into consideration, we propose using pretrained lightweight GNNs as representation extractors, and then detect anomalies with traditional machine learning (ML) detectors because of their lightweight characteristics (Fig. 2). In essence, our approach is not end-to-end; instead, we extract representations from pretrained GNNs. The primary computational cost is the one-time pretraining on a multi-action classification task, rather than training from scratch for each anomaly detection task.

We pretrain the lightweight GNNs based on [8], which uses self-distillation, a knowledge distillation (KD) model, to train and compress heavy models simultaneously. KD is a popular [5] compression method because it flexibly discovers lightweight models, both in terms of model size and model architecture. Self-distillation, specifically the BYOT [26] approach, extracts parts of the heavyweight model to form the compressed lightweight model without the need for careful redesign of the lightweight model architecture. During training, knowledge from the deeper layers is squeezed to guide the shallower layers, ensuring that the shallower layers achieve performance similar to that of the deeper layers.

Briefly, this paper focuses on anomaly detection using the extracted representations of actions by lightweight GNNs trained with BYOT. Our approach differs from end-to-end unsupervised anomaly detection by emphasizing reduced training costs while maintaining comparable detection performance. *To the best of our knowledge, we are the first to use representations from lightweight GNNs trained with BYOT on a multi-action recognition task and reuse them for anomaly detection with efficient traditional machine learning.*

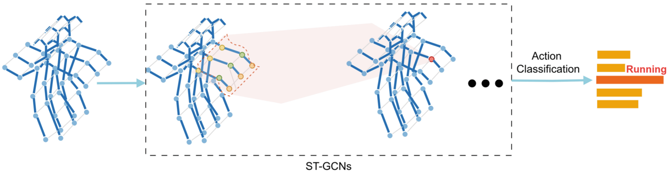


Fig. 1. The method ST-GCN [25]. Human joints are represented as blue dots, which are estimated from input videos. Each human skeleton is formed as a graph. The green node itself and its one-hop neighbors (yellow and pink nodes) are aggregated as the red node. The pink area represents the message-passing abstraction.

The following of this paper is organized as follows: Sect. 2 summarizes the previous works on GNNs, model compression and abnormal human action detections. Then, Sect. 3 depicts the methodology briefly, leading to Sect. 4 which evaluates the performance of our methods and analyzes the characteristics of the extracted representations. Section 5 provides the summary. Additionally, the terms “features” and “representations” are used interchangeably with the same meaning in this paper.

2 Previous Works

2.1 GNNs for Extracting Features

Transforming human action videos to graphs can protect personal privacy and reduce dimensionality. For instance, human skeleton graphs remove human faces and surrounding environments from original videos, thereby preventing the identification of observed subjects.

GNNs was initially introduced by Gori et al. [9] to extract the representations of graph data. They have since evolved into convolutional GNNs (ConvGNNs), which draw inspiration from convolutional neural networks (CNNs). ConvGNNs are further categorized into spectral ConvGNNs and spatial ConvGNNs. Spectral ConvGNNs process input data in spectral space, while spatial ConvGNNs convolve with k-order topological neighbors. ST-GCN belongs to the spatial ConvGNNs category.

GNNs for extracting action representations are classified as spatial-based approaches, spatiotemporal-based approaches and generated approaches [7]. ST-GCN is a typical instance of spatiotemporal-based approach that processes the spatial and temporal dimension inside one ST-GCN layer. Inspired by attention [24], AAGCN [22] adds spatial attention (nodes-level), temporal attention and channel attention to improve ST-GCN.

2.2 GNNs Compression

Common methods for obtaining a lightweight GNNs include simplifying GNN components or compressing GNNs [14]. The simplification is efficient but requires better knowledge for the specific GNN layers, while the GNNs compression can borrow the compression method of traditional neural networks (NNs). Among them, Knowledge Distillation (KD) is particularly popular because it allows robust performance without needing detailed knowledge of the model structure [5]. KD achieves this by distilling knowledge from a complex, heavyweight NN (teacher) to a simpler, lightweight NN (student).

There are typically three approaches to KD: offline distillation, which requires a pretrained heavy model; online distillation, which trains the lightweight NNs in real-time; and self-distillation, which simultaneously trains both heavy and light NNs [10]. Self-distillation can be regarded as a special online distillation, where the light NNs is a subset of the heavy NNs. It is more practical because it eliminates the need for carefully selecting architectures for both heavy and light NNs, enabling the rapid acquisition of a comparably lightweight NN. This paper adopts lightweight GNNs pretrained using the self-distillation framework proposed by [8].

2.3 Abnormal Action Detection

In addition to DL anomaly detectors, traditional unsupervised anomaly detectors typically include clustering methods, nearest neighbors methods, statistical methods, information theoretic methods and spectral methods [4]. The popular local outlier factor (LOF) [2] model relies on nearest neighbors and evaluates the anomalies by their densities. Gaussian Mixture Models (GMM) and Dirichlet Process Mixture Models (DPMM) [1] are statistical methods that use mixtures of parametric distributions, while the isolation forest (iForest) [13] employs isolation rules for anomaly detection. Other methods encompass ensemble methods like LODA, subspace methods such as SOD [23], and more.

To detect anomalies on the features in high dimensions, the common way is to directly work on the input high dimension features, which is simple and straightforward but may not always be optimal. Anomaly detection just requires distinguishing anomalies from other actions, rather than seperate each action. For instance, actions like staggering or falling differ significantly from normal daily activities in terms of speed and spatial movements, without needing to differentiate between various normal daily actions.

However, common representation extractors for human actions tend to emphasize detailed action features, which can overwhelm anomaly detection systems with excessive information. This not only biases anomaly detection but also consumes unnecessary time due to the high-dimensional nature of the data. One simple approach to mitigate this issue is feature reduction techniques such as PCA, t-Distributed Stochastic Neighbor Embedding (T-SNE) [16], which help remove redundant information and streamline anomaly detection processes.

3 Methodologies

3.1 Problem Definition

Regard the human skeleton in each frame as a graph $G = (\mathbf{V}, \mathbf{E})$, where the node set $\mathbf{V} = \{v_i; i = 1, \cdots, n\}$ is the skeleton joints and the edge set $\mathbf{E} = \{e_{ij}; 1 \leq i \leq n, 1 \leq j \leq n\}$ is the physical bone connections, the graph G's topology is further represented as an adjacency matrix $\mathbf{A}_{n\times n}$:

$$\mathbf{A}_{ij} = \begin{cases} 1, \text{if } e_{ij} \in \mathbf{E}, \\ 0, \text{else}, \end{cases} \tag{1}$$

with each item denoting whether there is an edge between the corresponding nodes. One example of the skeleton graph is shown in Fig. 1.

The features of nodes at each frame t are basically defined as the 3D coordinates $x_t^c = \{x_{t,1}^c, x_{t,2}^c, x_{t,3}^c\}$, with the numbers $1, 2, 3$ denoting three different axes, c indicates coordinates. To include the joint variations between frames t and $t+1$, we also take the one frame displacements as movements features, which is denoted as $x_t^m = \{x_{t+1,1}^c - x_{t,1}^c, x_{t+1,2}^c - x_{t,2}^c, x_{t+1,3}^c - x_{t,3}^c\}$.

Suppose the network f_l is the lightweight GCNs that has been trained on multi-actions recognition task, given the topology $\mathbf{A}$ and nodes feature $\boldsymbol{x}$ for one instance, the features $\boldsymbol{z}$ are extracted as

$$\boldsymbol{z} = f_l(\mathbf{A}, \boldsymbol{x}). \tag{2}$$

Dimensionality reduction techniques can be applied to $\boldsymbol{z}$ to further improve the results by removing redundant information. With a trained anomaly detector

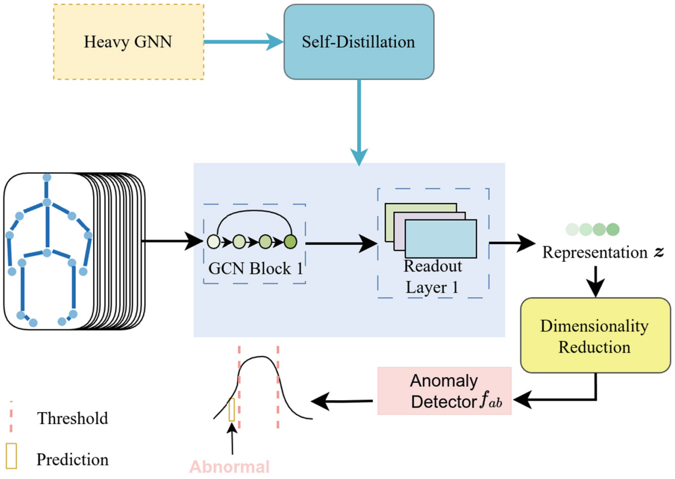


Fig. 2. The proposed anomaly detection framework based on self-distillation and traditional machine learning. The rectangles denote models while the rounded rectangles denote the processes. The light blue rectangle denotes the lightweight pretrained GCNs f_l, and the heavyweight GCNs f_h is denoted as the orange rectangle. When the prediction is smaller than the threshold, the input instance is classified as abnormal.

f_{ab}, given the threshold s_t, for one instance, its anomaly inference p is

$$p = \begin{cases} \text{Abnormal, if } f_{ab}(f_l(\mathbf{A}, \boldsymbol{x})) < s_t, \\ \text{Normal, else.} \end{cases} \tag{3}$$

The whole framework is shown in Fig. 2.

3.2 The Lightweight GCNs by Self-distillation

Self-distillation, specifically in this paper, BYOT [26], takes a subset of the heavy model f_h as the light model f_l. Examples of f_h, f_l are shown in Fig. 3, represented as rectangles in different colors and border lines. The f_h is the model colored in yellow, the light models $f_l^{b_1}, f_l^{b_2}, f_l^{b_3}$ are denoted by blue sparsed dashed border lines, purple dashed border lines, and green dotted border lines respectively. The denser a border line, the heavier the corresponding f_l. In our study the heavy model f_h can be either ST-CGN or AAGCN (see below).

The light compressed network f_l is solved by minimizing the distances between the shallow block and the deep block from the same backbone GCNs. Typical example pairs of the shallow and the deep block can be $(f_l^{b_1}, f_l^{b_3})$, $(f_l^{b_1}, f_h)$. The distances are defined as the combination of three components: the supervised action classification loss, the similarity distance between the predicted distributions of the shallow and deep block, and the feature similarity distance between the predicted representations of the shallow and deep block. For more details, please refer to [8].

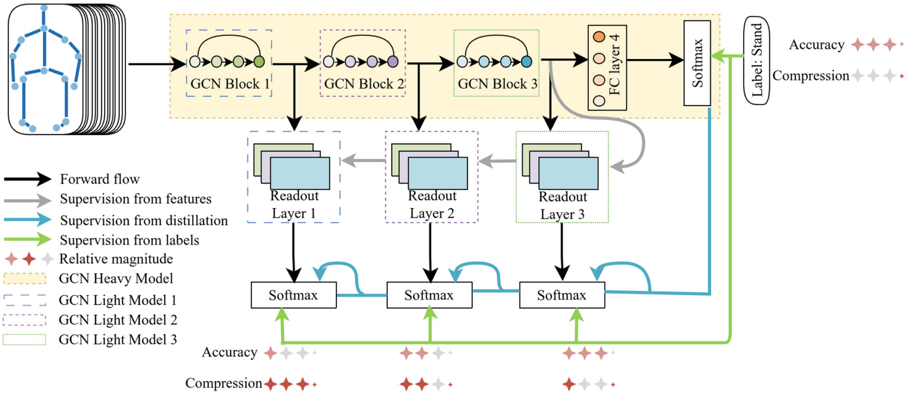


Fig. 3. The self-distillation compression framework, drawn based on [26], with a main difference at the GCN Block and input data. The GCN Block means one block of GCNs model. The FC means fully connection layer. Compression denotes the relative number of weights of each compressed model.

3.3 Unsupervised Anomaly Detection

Training Anomaly Detectors
Typical unsupervised anomaly detectors are trained only on normal actions, due to the difficulty of obtaining abnormal actions. The features $\boldsymbol{Z}$ of the whole dataset grabbed by f_l are in high dimensional space $\mathcal{R}^d$. Given $\boldsymbol{Z}$, suppose the anomaly detector is f_{ab}, if training is directly applied to $\boldsymbol{Z}$, then f_{ab} is solved by minimizing the losses defined by each f_{ab}.

However, as discussed in Sect. 2, the features $\boldsymbol{Z}$ in $\mathcal{R}^d$ contain excessive information for anomaly detection because their extractors are trained on multi-action classification tasks. To solve this, we propose using feature reduction instead of directly preforming anomaly detection on the original $\boldsymbol{Z}$. Typical feature reduction methods include PCA, T-SNE and Umap [18].

Anomaly Detectors
In this paper, the anomaly detectors DPMM, GMM, iForest and LOF are selected based on their computing cost and efficiency. One-class SVM is excluded due to its high memory requirements for large datasets. DPMM and GMM track distributions, while iForest compares tree heights and LOF focuses on density differences.

DPMM and GMM are mixture models that estimate the data distribution as a combination of multiple simpler distributions, such as Gaussian distributions for GMM. Each Gaussian distribution is considered as a component of the overall estimated distribution which is evaluated on the normal set only. Instances that fall outside the estimated distribution are classified as abnormal. GMM has a predefined fixed number of distribution components and uses maximum likelihood estimation (MLE) within the expectation-maximization (EM) framework [6], while DPMM has an adapative number of components that vary across datasets and is optimized within the Variational Inference (VI) framework. Specifically, DPMM draws priors from a Dirichlet Process, adopts Kullback–Leibler (KL) divergence as the measurement and at the end works on log marginal likelihood.

iForest is proposed based on tree models and focuses on global anomalies. One advantage of tree models is their good explainability. The basic assumption is that anomalies are rare and distinct from the normal set, and thus can be isolated earlier while generating a random forest. As the decision trees grow randomly, the features at each node are chosen randomly, and a random threshold is selected to divide the dataset in half. The dataset is continuously cutted away until all instances are isolated from one another. The average number of steps required to isolate normal instances is used as a reference. For new instances, anomalies are those that are isolated with fewer steps.

LOF, a density-based method, is popular for detecting local anomalies. For a given instance, the densities of its K-nearest neighbors are measured. If the instance's local density is significantly lower than that of its neighbors, it is marked as abnormal. The neighbors are identified using the K-nearest neighbor (KNN) algorithm.

4 Experiments

To evaluate the self-distilled GCNs for anomaly detection, we tested on NTU RGB+D [20] and Kinetics400 [11]. NTU RGB+D, built in 2016 using three Kinect V2 cameras in a lab setting, captures 60 actions across 56,880 videos. Each video includes depth map sequences, 3D skeletons, and infrared (IR) videos. The first 50 actions are single-subject actions, while the last ten are interaction actions. For simplicity, we refer to this dataset as NTU. Each skeleton consists of coordinates for 25 joints. The dataset is further split into NTU xsub and NTU xview based on cross-view and cross-subject splits. Kinetics400 collects videos from real-life scenarios and is more challenging compared to NTU RGB+D due to the presence of more subjects, partially occluded bodies, and complex environments. Kinetics400 captures 400 actions as RGB videos, and ST-GCN [25] processed it to skeleton joints with OpenPose [3], a robust tool that preserves 18 joints in 2D. GEPC [17], a popular paper on unsupervised anomaly detection using ST-GCN, selected the top 250 actions for pretraining to avoid extremely low performance due to difficult actions (e.g. incomplete skeletons, multiple subjects etc.). For clarity, we will refer to this subset as K250.

In total, 20 anomaly detection tasks are defined and tested, classified into speed-sensitive tasks, exercise tasks, and anomaly tasks related to daily activities. The task names are shown in Table 1. Most abnormal tasks preserve the same definitions as GEPC [17].

The GCN backbones selected in this paper are ST-GCN and AAGCN, which are fed with skeleton coordinates and movements features as input data (Sect. 3.1). The input data always includes two subjects. For the NTU dataset, since the first 50 actions are single-subject actions, the second subject is filled

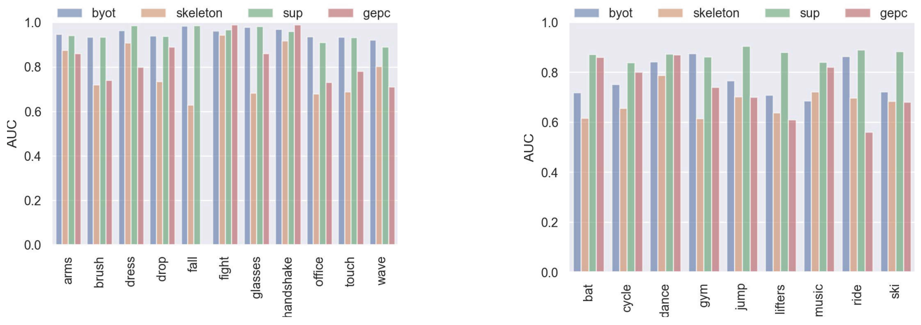
Table 1. The anomaly detection tasks for each category.

Category	Tasks
Speed-sensitive	Fall, fight, jump
Exercise	Bat, dance, gym, lifters, ride, ski
Daily activities	Arms, brush, dress, drop, glasses, handshake, office, touch, wave, cycle, music

with zeros. GEPC is used for comparison because it is also based on the ST-GCN backbone. However, GEPC uses input reconstruction for pretraining the ST-GCN extractor, which is unsupervised, whereas our extractors are trained using action labels. For a more fair comparison, we also compare the results released in [15], which uses ST-GCN for direct anomaly classification. It utilizes the URFall dataset [12] and their self-built dataset. Since their self-built dataset is not publicly available, we only compare results with URFall.

The evaluation score for anomaly detection is AUC, which calculates the area under the ROC curve. The closer the AUC is to 1, the better the performance of the corresponding anomaly detector. Unless specified otherwise, all the best AUC or other scores (recall, precision, f1) shown below are collected across all selected anomaly detectors.

4.1 Ablation Study



(a) NTU (b) K250

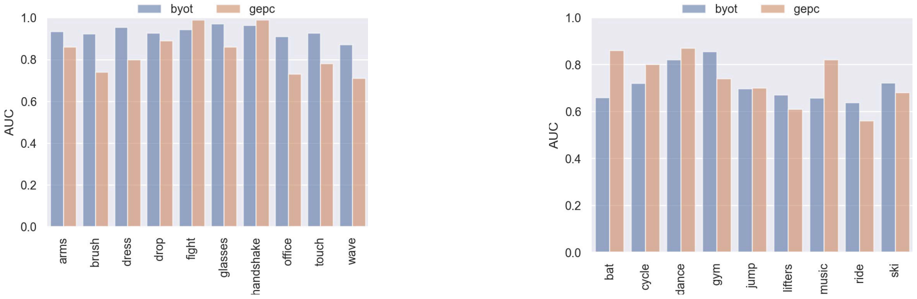
Fig. 4. The AUC for each task and dataset is presented, where ‘byot’ denotes the representations from backbones distilled through self-distillation, ‘sup’ stands for the representations from pure supervision (without self-distillation) using the same backbones, ‘skeleton’ indicates that the original skeleton input is directly fed into anomaly detectors, and ‘gepc’ marks the results released by GEPC [17]. All results from ‘byot’ and ‘sup’ are across ST-GCN and AAGCN, while ‘gepc’ is from ST-GCN based AutoEncoder. These results are the best obtained across all selected anomaly detectors: GMM, DPMM, iForest and LOF.

As shown in Fig. 4, except for the task Music, the AUC from raw skeletons always underperformed compared to the AUC from self-distillation (BYOT) across ST-GCN and AAGCN. This demonstrates that self-distillation effectively enhances the extracted representations to capture the semantic differences between normal and anomalous actions. These results are the best obtained across all selected anomaly detectors: GMM, DPMM, iForest and LOF.

4.2 Anomaly Detection Performance

Anomaly Detection Compared with SOTA

NTU and K250 Datasets. Figure 4 compares the performance of the representations from GEPC and those obtained through self-distillation across ST-GCN and AAGCN extractors. Except for tasks Fight and Handshake on NTU, tasks Bat, Cycle, Dance and Music on K250, features from BYOT consistently outperform those from GEPC in anomaly detection. For a fairer comparison, using the same ST-GCN and DPMM anomaly detector, our method still performs better than GEPC in most tasks (Fig. 5), on NTU (8 out of 10), and on K250 (5 out of 9). Notice that GPEC reconstructs the input for pretraining their ST-GCN based extractors, which is unsupervised, whereas our method leverages action labels. However, GEPC trained different extractors for different tasks, while our method only has a common extractor for all tasks of the dataset during anomaly detection. The difficulties of K250 causes our pretrained extractors to perform better only for easier tasks whose actions are simpler during pretraining.

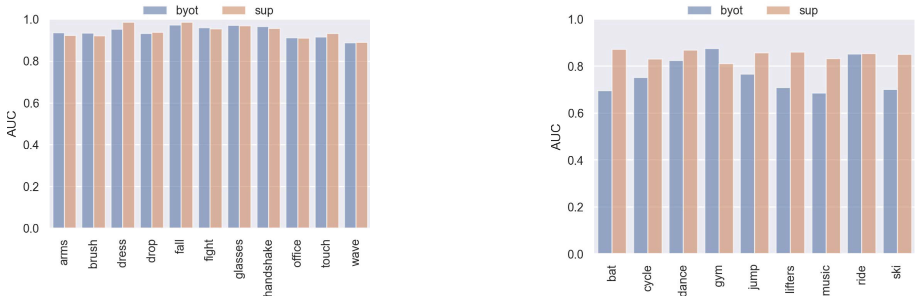


(a) NTU

(b) K250

Fig. 5. The AUC comparasion under the same setting of GEPC [17], with ST-GCN as the backbone and DPMM as the anomaly detector.

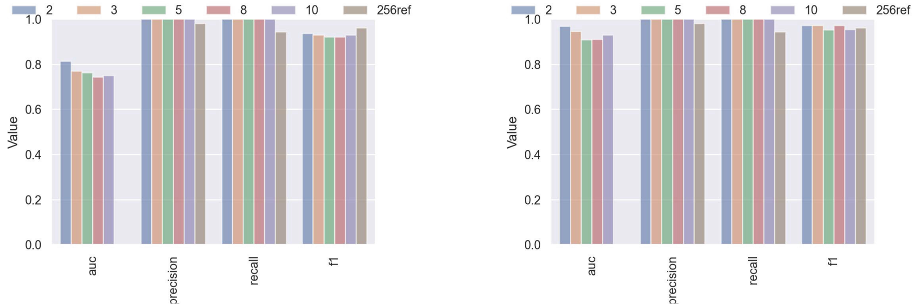
As shown in Fig. 6 which compares the performance across the lightest extractors ST-GCN8 and AAGCN8, on the NTU dataset, our method performs better or equal in 7 out of 11 tasks compared to representations from pure supervision (extractors pretrained by action labels without self-distillation across ST-GCN or



(a) NTU (b) K250

Fig. 6. The AUC comparison between the representations pretrained with self-distillation or without self-distillation across the lightest AAGCN8 and ST-GCN8 extractors, where 'byot' denotes the extractors trained with self-distillation, and 'sup' represents those without self-distillation.

AAGCN). For K250, only Gym outperforms the pure supervision approach, due to the difficulties of the dataset itself. K250 is larger and more challenging compared to NTU, leading to lower accuracy during pretraining by self-distillation and consequently poorer representations for anomaly detection. Nevertheless the lightweight GNN are still able to perform satisfactorily with the lightest extractors.

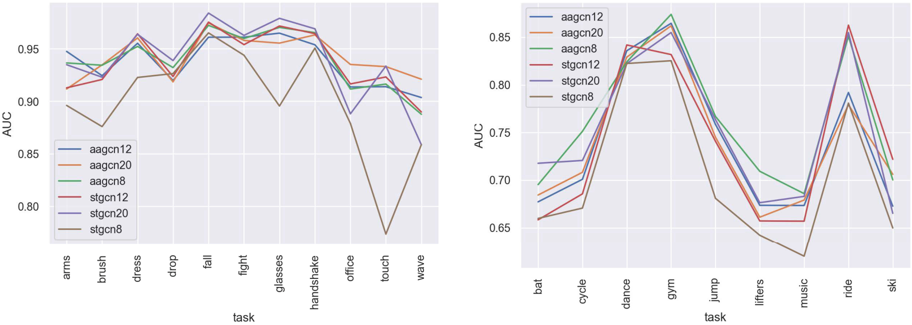


(a) ST-GCN (b) AAGCN8

Fig. 7. The best results of fall detection on URFall with numbers denoting different levels of feature reduction across PCA, T-SNE and Umap, where (a) summarizes the results across all ST-GCN extractors, and (b) collects the results of the lightest AAGCN8 extractor. The '256ref' represents the referenced results [15], which uses features in 256 dimensions and the ST-GCN20 extractor. Notice that AUC was not measured in [15].

URFall Dataset. Because of the unsupervised nature of GPEC, for a fairer comparison, we collected fall detection results on the URFall dataset. The URFall dataset contains 40 normal daily actions captured by camera 0 and 30 abnormal falling actions captured by cameras 0 and 1. During anomaly detection, 70% of the normal actions are used as the training set, while the remaining normal actions and all abnormal actions are used as the test set.

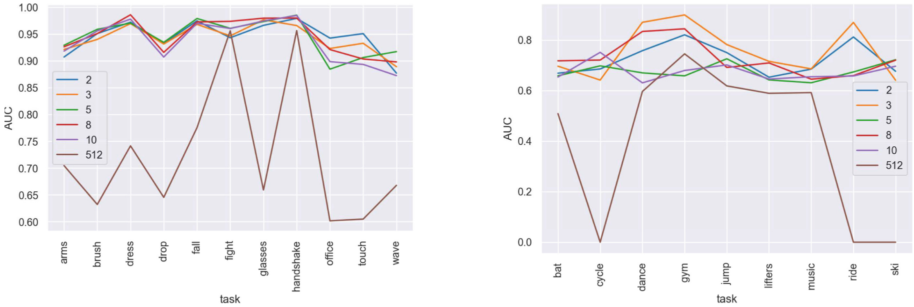
We compared our method with the supervised model proposed in [15], which extracts the features in 256 dimensions and measures performance with precision and recall. Considering the potential redundancy in features, we also feed anomaly detectors with features after reducing dimensionality using techniques PCA, T-SNE and Umap. The 256-dimensional features were reduced to 2, 3, and other dimensions. As demonstrated in Fig. 7a, our method outperforms [15] in precision and recall, even when the features are reduced to only 2 components, which is significantly fewer than the 256-dimensional features used in [15]. However, because of the small size of URFall and model structures, ST-GCN can easily overfit, while AAGCN is more steady during pretraining due to its attention mechanism. Figure 7b illustrates that AAGCN8 is better than the ST-GCN20 in [15] on URFall, and is also lighter.



(a) NTU (b) K250

Fig. 8. The AUC of each extractor for each task and dataset is presented. All extractors are compressed by self-distillation. The extractors, such as ‘stgcn20’ and ‘aagcn20’, represent the ST-GCN or AAGCN backbones at different compression levels. The smaller the number, the lighter the extractor.

Comparison of Lightweight GCN Extractors. Each GCN backbone structure follows the same design as in [8], where the heavyweight ST-GCN20/AAGCN20 are compressed to ST-GCN12/AAGCN12 or ST-GCN8/AAGCN8 during self-distillation. The ST-GCN8/AAGCN8 are the lightest, being compressed by approximately 60% compared to the ST-GCN20/AAGCN20. As shown in Fig. 8, the ST-GCN8 backbone fails to capture as much information compared to AAGCN8. The attention mechanism in AAGCN helps improve the extracted representations by adjusting attention weights during pretraining. In Fig. 8, AAGCN, especially AAGCN8, steadily stand out. AAGCN8 has a maximum AUC reduction of only 3.3% compared to AAGCN20 on the NTU dataset. On K250, AAGCN8 is better than AAGCN20 in most tasks, with a maximum difference of 7.2%. This might be due to the large capacity of AAGCN20 that could overfits on the difficult dataset K250.



(a) NTU

(b) K250

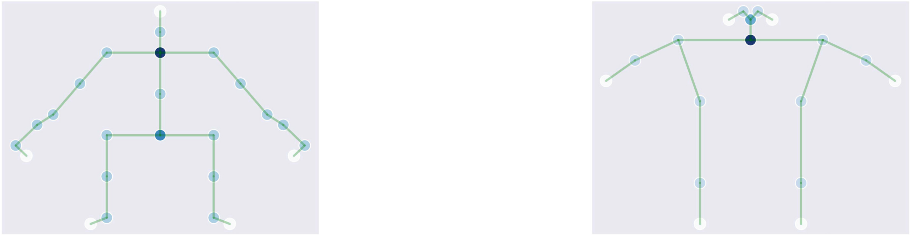
Fig. 9. AUC for different number of features for every task and dataset. All extractors are compressed by self-distillation. AUCs are collected across all anomaly detectors, while '512' is from iForest.

Improving the Representations for Anomaly Detection

Removing the Redundancy of Representations. To check the redundancy of representations for anomaly detection, we use dimensionality reduction techniques T-SNE, Umap and PCA to feed more disriminative features into anomaly detectors. As shown in Fig. 9, on both datasets, feature reduction always provides better results with some tasks prefering a few more features than others but far less than the original 512 features. The explaination is that the representations from self-distillation provides more information than anomaly detection needs, leading to a preference for fewer features. Within the same dataset, anomaly detection is simpler compared with multi-action classification (used for self-distillation), because it does not demand the detailed information to classify each action, instead, it only requires distinguishing between normal and abnormal actions.

Tasks that prefer more features usually require additional information from the hands or feet, which are nodes with lower degrees and therefore receive less information from other nodes compared to those with higher degrees. Adding more features helps gather information from nodes with higher degrees, aiding in detection. The relative degrees of each node in the skeleton graphs are shown in Fig. 10.

The Preference of Dimensionality Techniques of Each Task. Some tasks have their preferred feature reduction methods. As shown in Fig. 11, Umap achieves the best AUC in most tasks. Umap has advantages of preserving global structure



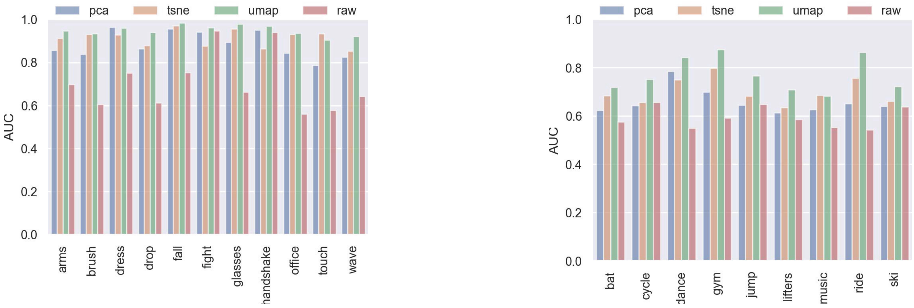
(a) NTU

(b) K250

Fig. 10. The degree of joint nodes for each type of skeleton graph is represented by color intensity: the darker the joint node, the higher its degree.

Table 2. Average improvement of AUC with different feature reduction methods across all tasks.

Dataset	PCA	T-SNE	Umap
NTU	17.94%	20.79%	24.45%
K250	6.52%	10.79%	17.69%



(a) NTU

(b) K250

Fig. 11. The best AUC under different feature reduction methods on each dataset, compared with the features without dimensionality reduction ('raw').

into the reduced set of features, while T-SNE is more interested in the local structure. PCA, on the other hand, tracks the axes with the largest variations and is linear.

For tasks Touch and Music, where the main action parts are the arms, T-SNE, which retains the local features of the arms, performs slightly better than Umap. Besides, for tasks Dress, Fight and Handshake in NTU and the task Gym in K250, PCA ourperforms T-SNE, which can be attributed to the quick movements along a specific direction as arms moving up and down or front and back, or legs moving up and down. As illustrated in Table 2, each feature reduction method consistently improves anomaly detection by at least 6.52% in AUC.

These results demonstrate that the proposed hybrid model is more beneficial compared to the unsupervised GEPC or the supervised method in [15]. First, the efficient feature extractor AAGCN8, is lighter compared with them. Second, GEPC requires training from scratch for each anomaly detection task, whereas our approach only needs to train the extractor once and then use it for all anomaly detection tasks.

5 Conclusion

Our goal was to evaluate if a lightweight GNN combined with a traditional anomaly detector could be effective on various tasks compared to other state-of-the-art methods. We have demonstrated that AAGCN8 is an accurate and efficient feature extractor for anomaly detection across 20 tasks. With feature reduction, compared to AAGCN20, AAGCN8 has a negligeable average AUC reduction of only 0.24% on NTU and an average improvement of 2.3% on K250, while being 60% lighter. Additionally, the AUC of our method with feature reduction is on average 11.4% higher than GEPC across all anomaly detection tasks on NTU and 3.25% higher on K250. Since the features extracted by our extractors are redundant for anomaly detection because they were trained on multi-action classification tasks, feature reduction techniques were also proposed to significantly improving the anomaly detection results.

However, our current extractors benefit from action labels. In the future, we aim to train the extractors in an unsupervised manner to capture representations that enhance anomaly detection.

Acknowledgements. The authors would like to thank Sébastien Roy for helping process URFall dataset.

References

1. Blei, D.M., Jordan, M.I.: Variational inference for dirichlet process mixtures (2006)
2. Breunig, M.M., Kriegel, H.P., Ng, R.T., Sander, J.: Lof: identifying density-based local outliers. In: Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data, pp. 93–104 (2000)
3. Cao, Z., Hidalgo Martinez, G., Simon, T., Wei, S., Sheikh, Y.A.: Openpose: real-time multi-person 2d pose estimation using part affinity fields. IEEE Trans. Pattern Anal. Mach. Intell. (2019)
4. Chandola, V., Banerjee, A., Kumar, V.: Anomaly detection: a survey. ACM Comput. Surv. **41**(3) (2009). https://doi.org/10.1145/1541880.1541882
5. Cheng, Y., Wang, D., Zhou, P., Zhang, T.: A survey of model compression and acceleration for deep neural networks (2020)
6. 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)

7. Feng, M., Meunier, J.: Skeleton graph-neural-network-based human action recognition: a survey. Sensors **22**(6) (2022). https://doi.org/10.3390/s22062091. https://www.mdpi.com/1424-8220/22/6/2091
8. Feng, M., Meunier, J.: A lightweight graph neural network algorithm for action recognition based on self-distillation. Algorithms **16**(12) (2023). https://doi.org/10.3390/a16120552. https://www.mdpi.com/1999-4893/16/12/552
9. Gori, M., Monfardini, G., Scarselli, F.: A new model for learning in graph domains. In: Proceedings. 2005 IEEE International Joint Conference on Neural Networks, vol. 2, pp. 729–734. IEEE (2005)
10. Gou, J., Yu, B., Maybank, S.J., Tao, D.: Knowledge distillation: a survey. Int. J. Comput. Vision **129**, 1789–1819 (2021)
11. Kay, W., et al.: The kinetics human action video dataset. arXiv preprint arXiv:1705.06950 (2017)
12. Kwolek, B., Kepski, M.: Human fall detection on embedded platform using depth maps and wireless accelerometer. Comput. Methods Programs Biomed. **117**(3), 489–501 (2014). https://doi.org/10.1016/j.cmpb.2014.09.005. https://www.sciencedirect.com/science/article/pii/S0169260714003447
13. Liu, F.T., Ting, K.M., Zhou, Z.H.: Isolation forest. In: 2008 Eighth IEEE International Conference on Data Mining, pp. 413–422. IEEE (2008)
14. Liu, X., et al.: Survey on graph neural network acceleration: an algorithmic perspective (2022)
15. Amsaprabhaa, M.: Multimodal spatiotemporal skeletal kinematic gait feature fusion for vision-based fall detection. Expert Syst. Appl. **212**, 118681 (2023). https://doi.org/10.1016/j.eswa.2022.118681. https://www.sciencedirect.com/science/article/pii/S095741742201716X
16. Van der Maaten, L., Hinton, G.: Visualizing data using t-sne. J. Mach. Learn. Res. **9**(11) (2008)
17. Markovitz, A., Sharir, G., Friedman, I., Zelnik-Manor, L., Avidan, S.: Graph embedded pose clustering for anomaly detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 10539–10547 (2020)
18. McInnes, L., Healy, J., Melville, J.: Umap: uniform manifold approximation and projection for dimension reduction. arXiv preprint arXiv:1802.03426 (2018)
19. Organization, W.H., Ageing, W.H.O., Unit, L.C.: WHO global report on falls prevention in older age. World Health Organization (2008)
20. Shahroudy, A., Liu, J., Ng, T.T., Wang, G.: Ntu rgb+ d: a large scale dataset for 3d human activity analysis. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1010–1019 (2016)
21. Shi, L., Zhang, Y., Cheng, J., Lu, H.: Two-stream adaptive graph convolutional networks for skeleton-based action recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 12026–12035
22. Shi, L., Zhang, Y., Cheng, J., Lu, H.: Skeleton-based action recognition with multi-stream adaptive graph convolutional networks. IEEE Trans. Image Process. **29**, 9532–9545 (2020)
23. Thudumu, S., Branch, P., Jin, J., Singh, J.: A comprehensive survey of anomaly detection techniques for high dimensional big data. J. Big Data **7**, 1–30 (2020)
24. Vaswani, A., et al.: Attention is all you need. In: Advances in Neural Information Processing Systems, vol. 30 (2017)
25. Yan, S., Xiong, Y., Lin, D.: Spatial temporal graph convolutional networks for skeleton-based action recognition. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 32 (2018). https://arxiv.org/abs/1801.07455

26. Zhang, L., Song, J., Gao, A., Chen, J., Bao, C., Ma, K.: Be your own teacher: improve the performance of convolutional neural networks via self distillation (2019)

Zero-Shot Spatio-Temporal Action Detection by Enhancing Context-Relation Capability of Vision-Language Models

Yasunori Babazaki(✉), Takashi Shibata, and Toru Takahashi

NEC Corporation, Kawasaki, Japan
{y_babazaki,takashi-shibata,t-taka}@nec.com

Abstract. We present a zero-shot spatio-temporal action detection framework that enhances the relational extraction capabilities of vision-language models. Zero-shot spatio-temporal action detection involves identifying a person's actions in a video and recognizing the time and place of these actions without prior training on those specific actions. Large-scale pre-trained vision-language models like CLIP exhibit zero-shot recognition capabilities for various tasks but struggle with extracting local features and relationships. By explicitly enhancing the extraction of person-context relationships in input videos and improving vision-language feature extraction, our proposed framework performs spatio-temporal action detection. It effectively captures local features and relationships between people and contexts while leveraging the strengths of zero-shot recognition from large-scale vision-language models. The two key components of our framework are person tracking in each input frame while ensuring smooth bounding-box shapes across frames, and the explicit interaction between visual features and language features in the shallow layers of visual feature extraction. We demonstrate the effectiveness of our framework through comprehensive experiments on two well-known action detection datasets, JHMDB and UCF101-24.

Keywords: Spatio-Temporal Action Detection · Vision-Language Model · Zero Shot

1 Introduction

Spatio-temporal action detection is a technique that recognizes a person's actions in a video and detects the time and place where these actions occur [1,15,20, 24,28]. By tracking the actions of each person chronologically throughout the video, this technique allows for a detailed understanding of the context relating to each individual and their environment, leading to more insightful analyses. This technique has a wide range of applications, including the detection of suspicious

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/978-3-031-78110-0_15.

A. Antonacopoulos et al. (Eds.): ICPR 2024, LNCS 15329, pp. 229–244, 2025.
https://doi.org/10.1007/978-3-031-78110-0_15

or unusual behavior in surveillance and the analysis of athletes' movements in sports. However, deep learning models generally require vast amounts of video data and high-cost annotations for each class, which greatly limits the scalability of action detection tasks.

Zero-Shot Spatio-Temporal Action Detection (ZSSTAD) aims to detect unseen actions without relying on large amounts of labeled training data [5]. Inspired by the new paradigm of pre-trained vision-language models like CLIP [16] and ALIGN [8], ZSSTAD leverages these models to detect unseen actions by linking text and image features. This approach enables the detection of new actions by pre-connecting textual and visual information.

On the other hand, research in action detection has demonstrated that capturing spatial and temporal interactions between individuals and objects within their context is crucial [3,15,20]. For instance, efforts have been made to develop mechanisms for efficiently extracting interactions between people and objects based on visual features [3,20]. These studies highlight the importance of accurately capturing spatial and temporal interactions in ZSSTAD while leveraging the strengths of zero-shot recognition models like CLIP.

In general, however, vision-language models represented by CLIP share the weakness that object-aware local feature extraction is inaccurate because it only aligns image features to the text features of alt-text that represents the entire image. Therefore, applying CLIP naively fails to accurately capture the spatial and temporal interactions with text feature that indicate actions (i.e., text prompts). Furthermore, existing state-of-the-art object tracking, e.g., ByteTrack [27], cannot adequately capture the relationship between objects and people due to the lack of bounding box (bbox) stability as shown in the top of Fig. 1.

In this paper, we propose a simple yet effective ZSSTAD framework that enhances the ability of vision-language models to extract person-context relationships in videos. The key innovation is the explicit enhancement of relational extraction between people and context in both input videos and vision-language feature extraction. Firstly, we perform person tracking in each input frame while ensuring that the bboxes for people maintain a smooth shape across frames. Because the bboxes representing both person and context are stable, the features extracted from these bboxes can effectively capture the relationships between the person and the context. Furthermore, inspired by recent advances in zero-shot object detection [9,13], we interact with text features at a shallow layer of visual feature extraction (see the bottom of Fig. 1). By introducing a relational extraction module that utilizes information from action prompts of the recognition targets, we improve the ability to extract person-context relationships that are tailored to the target actions. Our proposed framework offers a straightforward method to incorporate relational extraction in videos while leveraging the zero-shot capabilities of large-scale vision-language models and the strengths of previous action detection approaches. We demonstrate the effectiveness of our framework through experiments on two well-known action detection datasets, JHMDB [7] and UCF101-24 [18].

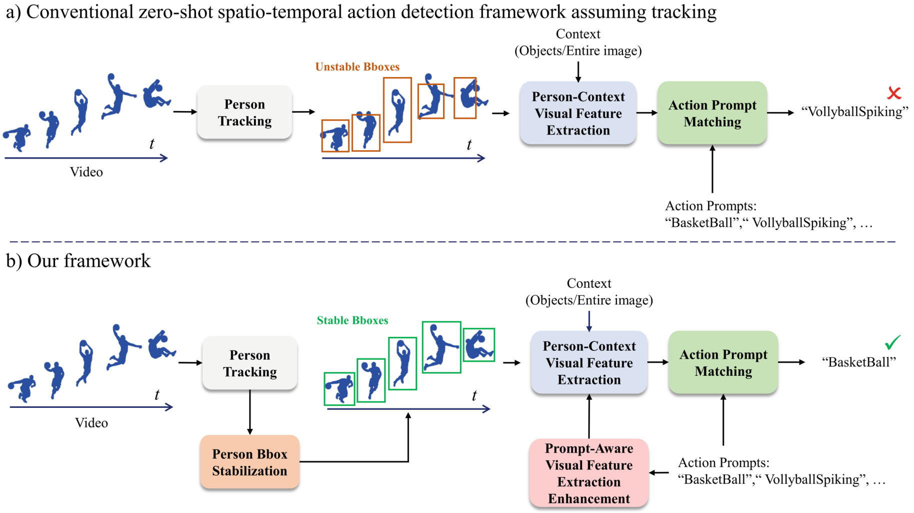

a) Conventional zero-shot spatio-temporal action detection framework assuming tracking
Video
t
Person Tracking
Unstable Bboxes
t
Context (Objects/Entire image)
Person-Context Visual Feature Extraction
Action Prompt Matching
"VollyballSpiking"
Action Prompts: "BasketBall"," VollyballSpiking", …
b) Our framework
Video
t
Person Tracking
Person Bbox Stabilization
Stable Bboxes
t
Context (Objects/Entire image)
Person-Context Visual Feature Extraction
Prompt-Aware Visual Feature Extraction Enhancement
Action Prompt Matching
"BasketBall"
✓
Action Prompts: "BasketBall"," VollyballSpiking", …

Fig. 1. Overviews of conventional zero-shot spatio-temporal action detection framework assuming tracking and our proposed framework.

Our contributions are as follows:

- We propose a zero-shot spatio-temporal action detection framework which enhances the relational extraction capability of vision-language models.
- We introduce person tracking that incorporates pose estimation to ensure smooth bbox shapes between frames for action detection, along with a module that enhances relationship extraction capabilities using text information (i.e., action prompts).
- We demonstrate the effectiveness of our framework through experiments on two well-known action detection datasets, JHMDB and UCF101-24.

2 Related Work

Vision-Language Models for Image Recognition. The field of Vision-Language Models (VLM) focuses on integrating image and text information to enhance mutual understanding. This area has numerous applications, including image captioning, image retrieval, and visual question answering. Notably, foundational models in VLM, such as CLIP [16], ALIGN [8], and their variants [2,12], which learn integrated visual and language representations, have achieved remarkable performance in zero-shot image classification. These models have become fundamental methods in the field of recent image and video recognition. In this paper, we leverage these VLMs to achieve ZSSTAD.

Vision-Language Models for Video Recognition. Inspired by the success of CLIP, which learns to associate language and images and enables image classification without the need for labeled training data, several recent studies have proposed extending CLIP to video data for video recognition [6,14,23,26]. These methods adapt CLIP to video data using various approaches, facilitating zero-shot video recognition. For instance, ActionCLIP [23] integrates a temporal feature aggregation layer on top of the image encoder to model temporal dynamics, while X-CLIP [14] introduces cross-frame attention for temporal modeling. CLIP-VIP [26] presents a more efficient approach to applying CLIP in the temporal domain, achieving high zero-shot recognition performance. While these studies primarily focus on zero-shot recognition across entire videos, our research centers on ZSSTAD. Specifically, we aim to identify the spatio-temporal positions of individuals within videos and perform zero-shot action recognition.

Recently, spatio-temporal action detection has begun to address zero-shot scenarios. To the best of our knowledge, the only existing ZSSTAD method, iCLIP [5], achieves this by utilizing pre-trained vision-language models and proposing interaction modules that extract relationships between people and content in videos. These modules also refine text features by considering image features, enabling ZSSTAD through the alignment of visual and text features for accurate action classification in videos. However, the integration of person tracking, which is essential for practical applications, has not been studied. Additionally, the exploration of relationship extraction between visual features based on text features (i.e., action prompts) remains unexplored, leaving room for further research.

Spatio-Temporal Action Detection. Action detection, which aims to spatio-temporally localize individuals and recognize their actions in videos, is a key task that advances video understanding. This task is gaining attention due to its diverse applications, such as activity monitoring and abnormal behavior detection. Action detection has seen significant advancements through deep learning methods, which generally fall into two main approaches: (i) a two-stage framework that first performs independent person/object instance detection followed by action recognition [15,20], and (ii) an end-to-end framework which conducts instance detection/action recognition in an end-to-end manner [1,24,28]. In this paper, we achieve zero-shot spatio-temporal action detection based on the two-stage framework, which facilitates the utilization of off-the-shelf person trackers, detectors, and CLIP.

3 Proposed Method

3.1 Overview

Our proposed method aims to improve the performance of ZSSTAD by enhancing the ability to extract spatio-temporal person-context relationships under the unstable conditions of person tracking typically encountered in real-world applications. We achieve this by introducing two novel mechanisms: 1) the Person BBox Stabilizer, which stabilizes the bboxes obtained from person tracking, and

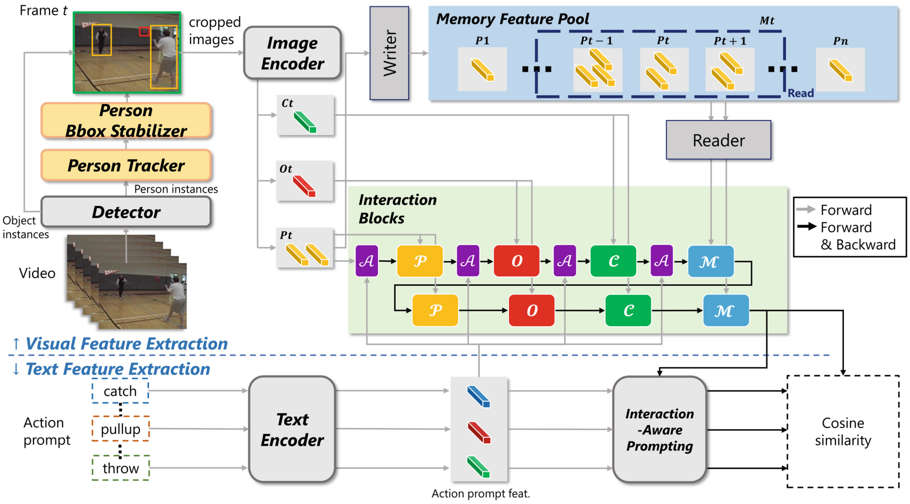


Fig. 2. Overview of our proposed framework: The figure above the dashed line illustrates the detector for people and objects, the person tracker, and the person bbox stabilizer. Using an image encoder, the framework extracts features for the person, object, and context (whole image) at frame t. Interaction blocks indicate that visual features for action recognition are derived by feeding these features into the action-prompt-aware interaction ($\mathcal{A}$), person-person interaction ($\mathcal{P}$), person-object interaction ($\mathcal{O}$), person-context interaction ($\mathcal{C}$), and memory feature interaction ($\mathcal{M}$). The memory features, which are time series of person features used in $\mathcal{M}$, are written to and read from the Memory Feature Pool. The figure below the dashed line shows that the text encoder computes the text embedding for each action label. We obtain the text embedding through Interaction-Aware Prompting to combine the text embedding with the interaction feature and calculate the similarity between visual and text embeddings.

2) the Action-Prompt-Aware Interaction Block, which aids in modeling relationships pertinent to the target actions. The proposed framework consists of two main components: visual feature extraction and text feature extraction, as illustrated in Fig. 2. In the following sections, we provide detailed descriptions of our ZSSTAD framework and the aforementioned mechanisms.

For the visual feature extraction, we first separate a video into consecutive frames, $\mathbf{F} = [F_1, F_2, ..., F_t, ...F_N]$, where N represents the total number of frames in the video. The detector extracts person/object instances as bboxes from each input frame, followed by the person tracker. The tracking results are fed into the person bbox stabilizer to obtain stable bboxes, which are suitable for action recognition. Person instance-level features, $P_t \in \mathrm{R}^{N_{Pt} \times D}$, and object instance-level features, $O_t \in \mathrm{R}^{N_{Ot} \times D}$, are computed from the image by cropping the person/object regions of interest with the bboxes using the pre-trained CLIP image encoder. Here, N_{Pt} and N_{Ot} are the number of person and object bboxes detected in t-th frame respectively, and D is the feature dimension. Additionally, the context feature, $C_t \in \mathrm{R}^{1 \times D}$, is extracted from the entire image using the same image encoder. P_t is stored in the Memory Feature Pool for temporal interaction modeling. To model multi-type interactions that aid in action recognition, we introduce interaction blocks, similar to those in [3,5,20]. Given different P_t, O_t, C_t, and sets of person memory features M_t read from the Memory Feature Pool, the interaction blocks output interaction features $f_{A_t} = IB(P_t, O_t, M_t, \phi_i)$, where ϕ_i represents the parameters in the interaction blocks. f_{A_t} is then passed to the consine-similarity-based action matching for final predictions.

The interaction blocks consist of person-person interaction, $\mathcal{P}$, person-object interaction, $\mathcal{O}$, person-context interaction, $\mathcal{C}$, and memory interaction blocks, $\mathcal{M}$, as presented in [3,5,20], along with the action-prompt-aware interaction, $\mathcal{A}$, proposed in this work. We adopt the order of $\mathcal{P}$, $\mathcal{O}$, $\mathcal{C}$ and $\mathcal{M}$ units as demonstrated in [3]. These interaction blocks provide richer representations for distinguishing action recognition.

For the text feature extraction, we input each action label name into CLIP's pre-trained text encoder to obtain the text features. These text features are then fed into $\mathcal{A}$ to achieve precise alignment between visual and language features, as proposed in [9,13]. Additionally, the interaction-aware prompting [5] refines the text features using f_{A_t}, making them suitable for action matching and output final text features, $f_{T_t} \in \mathrm{R}^{Na \times D}$ (Na is the number of target actions), for action classification. In the training phase, the image and text encoders are frozen, and the remaining parameters along the bold black arrows shown in Fig. 2 are updated by computing cosine similarity between f_{A_t} and f_{T_t}, and using the loss function identical to Eq.(5) in [5].

3.2 Person Bbox Stabilizer

Visual feature extraction can identify features that are more suitable for consistent action recognition when the input images have stable sizes and positions. If the bboxes fluctuate significantly, the feature extraction process must handle images with varying scales and positions, which complicates accurate action recognition. This challenge is particularly pronounced in zero-shot settings, where it is impossible to train models for each target action, and bbox instability cannot be directly addressed. To address this issue, we introduce a mechanism for stabilizing person bboxes, known as the Person Bbox Stabilizer (PBS).

Pose estimation is a technique used to precisely extract the position of individuals in an image. By employing pose estimation to re-localize the tracking bbox, a more stable bbox can be achieved. In our approach, as outlined below, we apply top-down pose estimation [19] within the bbox of the i-th individual at frame t, denoted as B_t^i, obtained through tracking.

$$\mathcal{J}_t^i = \{(X_{t,k}^i, Y_{t,k}^i)\}_{k=1}^K = \psi^P(\phi^{Crop}(F_t; B_t^i)), \tag{1}$$

where ψ^P and ϕ^{Crop} represent the pose estimation and image cropping operations with bboxes, respectively. Here, $\mathcal{J}_t^i = \{(X_{t,k}^i, Y_{t,k}^i)\}_{k=1}^K$ is the set of the xy coordinates $(X^i_{t,k}, Y^i_{t,k})$ of each joint, where k and K is the joint index of each bbox B^i_t. The pose-estimation-based stabilized bbox $\hat{B}^i_t = (x^{min}, x^{max}, y^{min}, y^{max})$ is formally given by

$$\hat{B}^i_t = \begin{pmatrix} x^{min} \\ x^{max} \\ y^{min} \\ y^{max} \end{pmatrix} = \begin{pmatrix} \min_k \{X^i_{t,k}\}^K_{k=1} & - & \frac{w}{2} \cdot r \\ \max_k \{X^i_{t,k}\}^K_{k=1} & + & \frac{w}{2} \cdot r \\ \min_k \{Y^i_{t,k}\}^K_{k=1} & - & \frac{h}{2} \cdot r \\ \max_k \{Y^i_{t,k}\}^K_{k=1} & + & \frac{h}{2} \cdot r \end{pmatrix}, \tag{2}$$

$$\begin{pmatrix} w \\ h \end{pmatrix} = \begin{pmatrix} \max_k \{X^i_{t,k}\}^K_{k=1} - \min_k \{X^i_{t,k}\}^K_{k=1} \\ \max_k \{Y^i_{t,k}\}^K_{k=1} - \min_k \{Y^i_{t,k}\}^K_{k=1} \end{pmatrix}, \tag{3}$$

where r is a scaling parameter, and we set this parameter to 0.4 in this paper, referencing the training data, to sufficiently encompass the person. The sensitivity of the performance to r is discussed in the supplementary material. The process of calculating the bboxes based on these poses is expected to stabilize the bboxes.

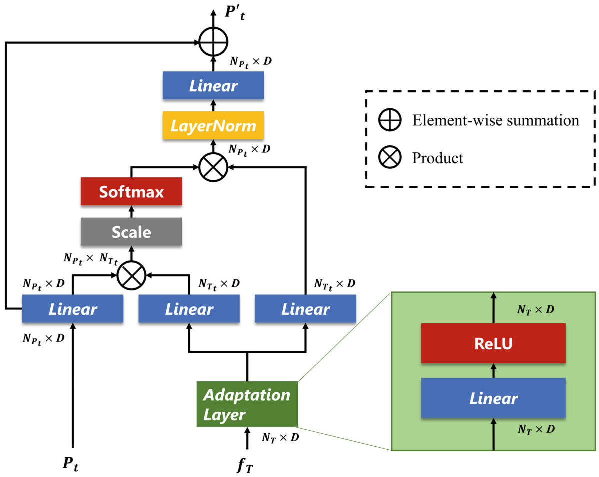


Fig. 3. Architecture of the Action-Prompt-Aware Interaction Block. The query input is the feature of the target person, and the key/value input consists of the text features passed through the adaptation layer.

3.3 Action-Prompt-Aware Interaction Block

In a zero-shot setting, there is no predefined limit on the number of action classes, necessitating the capture of diverse behavioral patterns. Understanding how each action relates to other contexts, such as objects and people, becomes crucial. To enhance the capability of relationship extraction between a person and other contexts, we introduce an Action-Prompt-Aware Interaction Block (APAIB), which enriches action features, f_{A_t}, by incorporating a relational extraction mechanism for the action prompts of the recognition target. This improves the ability to extract relationships between individuals and the context relevant to the target.

Recent studies on zero-shot object detection (ZSOD) have reported that early fusion between visual and language features is essential for image understanding [9,13]. Inspired by these recent ZSOD studies, APAIB has been designed. As shown in Fig. 3, APAIB employs a transformer-like attention mechanism [21], as adopted in [3,5,20], where the feature of the target person is used as the query, and the text features are used as the key/value, to extract the relationships between the person feature and the text features. By considering the information from the action prompts in the shallow layers using this block, it is possible to facilitate the extraction of relationships in visual feature extraction tailored to the recognition target actions.

Unlike the attention mechanisms in [3,5,20], APAIB introduces the following adaptation layer consisting of a learnable non-linear layer that transforms text features:

$$f_T{}' = \mathrm{ReLU}(Linear(f_T)). \tag{4}$$

This adaptation layer converts text features to better align with visual features, resulting in improved extraction of relationships and producing more accurate action features.

4 Experimental Results

4.1 Experimental Settings

Dataset. The JHMDB dataset [7] contains 21 types of action classes and 928 videos. Each class includes up to 55 clips, with each video clip containing a single action. A total of 31,838 frames are annotated, where bboxes, tracking tubelets, and actions are annotated for each person instance in each frame. The performance is evaluated using frame mean Average Precision (mAP) on split 1 of the dataset. An Intersection over Union (IoU) threshold for frame mAP is 0.5. On the other hand, the UCF101-24 [18] dataset is a subset of the UCF101 dataset, specialized for spatio-temporal action detection. It consists of 24 action categories and 3,207 untrimmed videos. Person bboxes with action labels are annotated frame by frame. Unlike JHMDB, UCF101-24 contains scenes where multiple people appear in a single video frame, often with significant overlap between individuals. We evaluate our method on the first split of this dataset and report frame mAP with an IoU threshold of 0.5.

Implementation Details. We use ByteTrack [27] with YOLOX [4] as the person tracker. For object detection, Faster-RCNN [17] trained on MSCOCO [11] is used as our object detector. ResNet-50-FPN and ResNet-101-FPN [10, 25] are used as the backbone of the Faster-RCNN for JHMDB and UCF101-24, respectively. We apply HRNet [22] pre trained with MSCOCO to obtain the pose from each person bbox for the PBS. For the image encoder and the text encoder, we use pre-trained CLIP [16] model ViT-B/16.

Training and Inference. In our experiments, we use 32 consecutive frames as network input. During the training phase, ground truth person bboxes and detected object bboxes are used to extract P_t and O_t. During inference, we use detected person and object bboxes to obtain the results. We freeze the image and text encoders of CLIP and train the network with the SGD optimizer, following the backward path indicated by bold black arrows in Fig. 2. For JHMDB, the batch size is set to 8 and the base learning rate is 2.0e–4. We train the network for 7k iterations, using a linear warm-up scheduler for the first 0.7k iterations. For UCF101-24, the batch size is set to 8 and the base learning rate is 2.0e–4. We train the network for 14k iterations, applying the linear warm-up scheduler for the first 0.14k iterations.

Table 1. Main results on JHMDB and UCF101-24. In this table, we display our best mAP results which include all proposed components. In *Oracle* and *Detection*, *Oracle* denotes that the input person bboxes are ground truth, while *Detection* indicates predictions. When using *Oracle*, the PBS of our method is not applied. The results for two training/evaluation split patterns of the action classes are shown: 1) 75%:25% class split and 2) 50%:50% class split. The best results are shown in **bold**.

	Method	JHMDB		UCF101-24	
		75%:25%	50%:50%	75%:25%	50%:50%
Oracle	Baseline (iCLIP [5])	69.6	45.5	90.3	66.1
	Ours	**70.4** (+0.8)	**45.9** (+0.4)	**92.1** (+1.8)	**66.4** (+0.3)
Detection	Baseline (iCLIP [5])	61.2	42.5	28.8	17.9
	Ours	**64.2** (+3.0)	**42.9** (+0.4)	**33.5** (+4.7)	**21.0** (+3.1)

4.2 Main Results

To evaluate the performance of our framework in a zero-shot scenario, we conducted experiments using datasets of unseen action classes that were not included during training, ensuring that the training and evaluation classes were mutually exclusive. We adopted two experimental setups with different action class splits: 1) 75%:25% and 2) 50%:50%. The 75%:25% class split follows the division provided in [5], while the 50%:50% split uses random sampling to divide the action classes. We compared the effectiveness of our framework against iCLIP [5], which served as the baseline. For a fair comparison, we evaluated performance using person bboxes provided by the person tracker employed in this paper and object bboxes from the object detector, consistent with the setup used in iCLIP. iCLIP was trained similarly to our framework, with the CLIP image and text encoders kept frozen.

Table 1 presents the evaluation results for JHMDB and UCF101-24. We report results for both settings: one utilizing ground truth person bboxes as person instances (*Oracle*) and the other using predicted person bboxes (*Detection*). The PBS is not applied in the *Oracle* setting. In the 75%:25% class split under the *Oracle* setting, our model outperformed the baseline by +0.8 mAP on the JHMDB and +1.8 mAP on the UCF101-24. Furthermore, in the *Detection* setting, our model showed improvements of +3.0 mAP on JHMDB and +4.7 mAP on UCF101-24 compared to the baseline. These improvements are more substantial than those observed in the *Oracle* setting, indicating that the proposed method contributes more effectively to performance enhancement in realistic scenarios where detection results may be unstable.[1] In the 50%:50% class split, we observed consistent improvements similar to those in the 75%:25% split for both datasets.

4.3 Ablation Study and Analysis

We conducted ablation experiments and analyses using the 75%:25% class split to assess the effectiveness of each key mechanism in our framework, specifically APAIB and PBS. Following this, we present a qualitative analysis.

Impacts of Two Key Mechanisms. We evaluated the effectiveness of APAIB and PBS using the JHMDB. Table 2 shows the results of the ablation experiments. It was observed that both APAIB and PBS individually contributed to the performance improvement. We found that our framework, which combines both APAIB and PBS, achieved the highest performance.

APAIB: Importance of the Adaption Layer and Insertion Point. We conducted ablation experiments focusing on APAIB using the JHMDB dataset. First, we investigated the importance of the adaptation layer, which is distinct from other interaction blocks. Table 3 presents evaluation results with and without the adaptation layer. The APAIB with the adaptation layer outperformed the version without it. Although text features are trained to align with visual features, they are not necessarily optimized for extracting relationships among

[1] We note that our results exhibit lower performance compared to the mAP reported in the original paper [5] due to issues with tracking matching, resulting in cases where tracked person instances are not output. In particular, UCF101-24 includes scenes with multiple individuals and frequent occlusions, which pose challenges for person tracking and consequently lead to reduced performance.

Table 2. Impacts of APAIB and PBS on JHMDB. The best result is shown in **bold**.

Method	Mechanism		mAP
	APAIB	PBS	
Baseline (iCLIP)			61.2
w/APAIB	✓		63.3 (+2.1)
w/PBS		✓	62.7 (+1.5)
Ours	✓	✓	**64.2** (+3.0)

persons and contexts in the image. Our results show that incorporating an adaptation layer with non-linear transformation can better tailor the text features for relationship extraction.

We then experimented with different insertion points for APAIBs within the interaction blocks. The interaction blocks, excluding APAIBs, consist of $\mathcal{P} \rightarrow \mathcal{O} \rightarrow \mathcal{C} \rightarrow \mathcal{M} \times 2$ stages, which have been reported as the best unit order in [5]. We investigated the effect of three different APAIB insertion patterns: 1) Before each interaction block in the first stage, $\mathcal{A} \rightarrow \mathcal{P} \rightarrow \mathcal{A} \rightarrow \mathcal{O} \rightarrow \mathcal{A} \rightarrow \mathcal{C} \rightarrow \mathcal{A} \rightarrow \mathcal{M}$, denoted as "1st stage", 2) Before each interaction block in the second stage, following the same setting as the first stage, denoted as "2nd stage", 3) Before each interaction block in all stages, denoted as "all stages". Table 4 shows the results for each pattern. We found that inserting APAIBs before each interaction block in the first stage achieved the best performance. This suggests that early-stage fusion of visual and language features is beneficial for zero-shot action recognition, consistent with findings in zero-shot object detection [9,13].

Performance of Each Action Class. To enable a more detailed comparison of the recognition results, we present the average precision for each class not used during training. Table 5 shows the results for JHMDB. In the *Oracle* setting, where ground truth person bboxes were used without the PBS, higher performance was achieved for half of the action classes. Additionally, in the *Detection* setting, which is closer to real-world applications using predicted bboxes, the

Table 3. Comparison of results with and without APAIB's adaptation layer on the JHMDB. The best result is shown in **bold**.

	mAP
w/o Adaptation Layer	61.2
w/ Adaptation Layer	**63.3** (+2.1)

Table 4. Insertion position of APAIB in interaction blocks. The experiments are conducted on JHMDB. The best result is shown in **bold**.

Insertion Position	mAP
1st stage	**63.3**
2nd stage	60.3
All stages	63.2

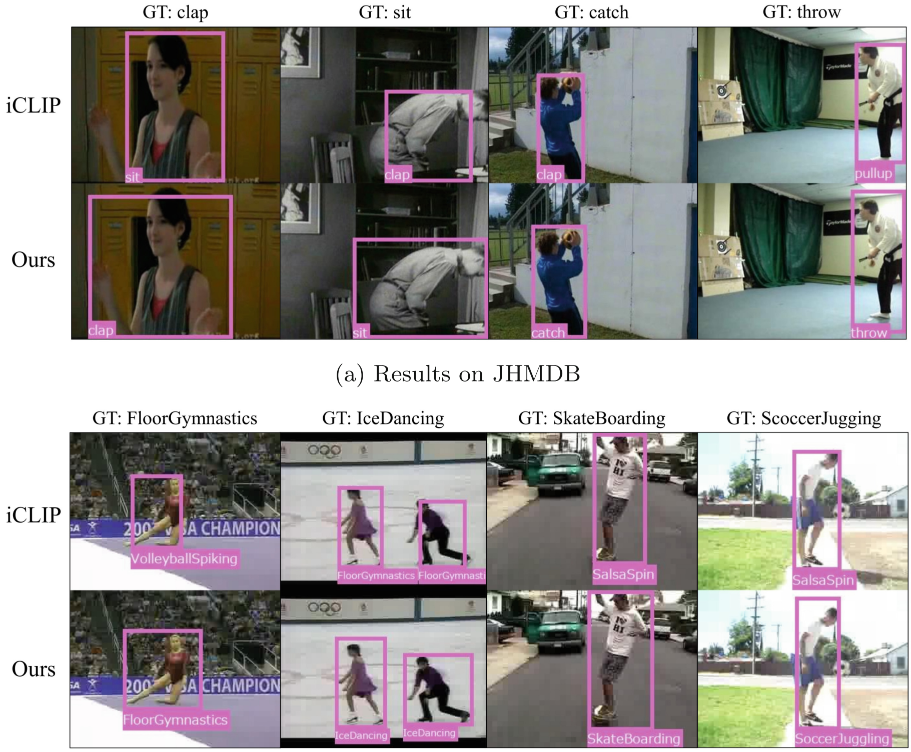
Table 5. Performance of each action class on the JHMDB. The best results are shown in **bold**.

	Method	Action Class						mAP
		catch	clap	pullup	sit	throw	wave	
Oracle	Baseline (iCLIP [5])	77.2	79.2	**100**	**67.3**	37.1	**57.1**	69.6
	Ours	**82.0**	**79.9**	**100**	63.2	**44.8**	52.5	**70.4** (+0.8)
Detection	Baseline (iCLIP [5])	71.3	73.2	99.9	42.1	36.8	**43.9**	61.2
	Ours	**77.5**	**75.7**	**100**	**42.5**	**47.7**	41.9	**64.2** (+3.0)

introduction of both APAIB and PBS led to improved performance for most action classes compared to the baseline, with a maximum improvement of +10.9 mAP. Due to space limitations, please refer to the supplementary material for the results of each class in UCF101-24.

Visualization: Qualitative Analysis. We performed a qualitative analysis based on the visualization of recognition results. Figure 4 shows examples of recognition outcomes on JHMDB and UCF101-24. In action recognition, person poses serve as crucial cues. For instance, in the cases of "clap" and "sit" on JHMDB, the proposed method successfully identifies the actions due to the bboxes that encompass person poses, as enabled by the PBS, shown in Fig. 4(a). In contrast, for actions like "catch" and "throw," although the bboxes cover the poses, the existing method fail to correctly identify the actions. For these actions, objects such as a ball and targets also play a significant role in the recognition process. These examples suggest that the successful action recognition achieved by our method is largely due to the introduction of the APAIB, which enhances the model's ability to understand the relationship between individuals and their context within the video. The superior performance of the proposed method is similarly evident in the results for UCF101-24, as shown in Fig. 4(b).

Discussion: Why is the Proposed Stabilizer Effective? To experimentally elucidate the factors contributing to the performance improvement brought about by the PBS, we further investigated the properties of the bboxes after the introduction of the PBS, based on the "Baseline w/APAIB". Figure 5 illustrates the distribution of correct and incorrect samples relative to the deviation of the bbox from the ground truth before and after the introduction of the PBS. To clearly demonstrate the effects of the PBS and facilitate trend analysis, we focus on the action class "clap", which has shown the greatest improvement, as indicated in the supplementary material. We display samples that are matched with ground truth bboxes at an IoU threshold of 0.5. The vertical axis represents the offset of the center point, which normalizes the difference between the centers of the ground truth bbox and the predicted bbox by the height of the ground truth bbox. The horizontal axis shows the H ratio and W ratio, which are the ratios of



(a) Results on JHMDB

(b) Results on UCF101-24

Fig. 4. Visualization for quantitative analysis on JHMDB and UCF101-24. The top and bottom rows show the recognition results of iCLIP and our method, respectively. Each figure shows the ground truth class labels at the top.

the predicted bbox width (W_{pred}) and height (H_{pred}) to the ground truth bbox width (W_{gt}) and height (H_{gt}), respectively. These are defined as W_{pred}/W_{gt} and H_{pred}/H_{gt}. Samples were classified as correct or incorrect based on the top-1 action class of each sample.

Focusing on the offset of the center point, we observed that both correct and incorrect samples showed greater spread after the introduction of the PBS, indicating that this parameter did not significantly affect accuracy. Additionally, examining the W ratio without the PBS revealed that incorrect samples were predominantly found in regions where the predicted bbox was significantly larger than the ground truth (W ratio > 1.5). In contrast, such samples were reduced with the introduction of the PBS. Furthermore, the use of the PBS increased the number of samples where the bbox was slightly larger than the ground truth ($1.0 < W$ ratio < 1.5). Considering the performance improvement with the PBS, this suggests that a bbox slightly larger than the ground truth, which encompasses the person more broadly, is advantageous for recognition compared

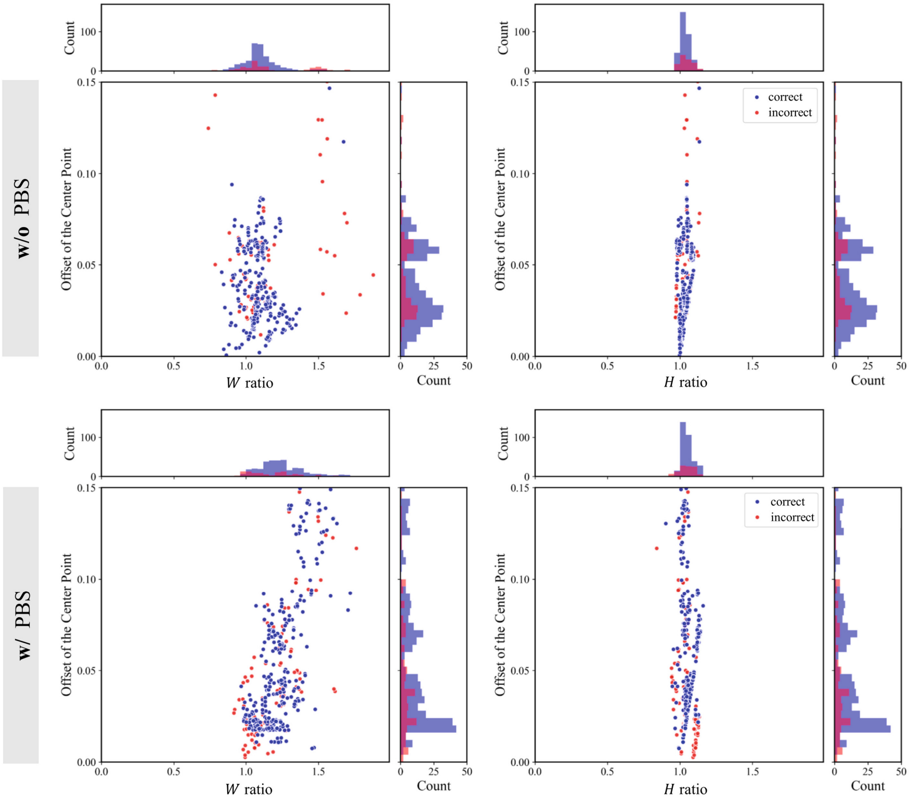


Fig. 5. Distributions of correct and incorrect recognitions in relation to the deviation of the bbox from the ground truth before and after introducing the PBS. The distributions are evaluated by using the "clap" action class on JHMDB.

to a tighter bbox around the person's contour. Note that the H ratio showed little variation before and after the incorporation of the PBS, with no significant trend differences observed.

5 Conclusion

We proposed a ZSSTAD framework that enhances the relationship extraction capabilities of vision-language models, assuming person tracking, which is crucial for practical applications. The key mechanisms of our framework are: 1) the PBS, which stabilizes the bboxes obtained from person tracking and aids in accurately extracting action features, and 2) the APAIB, which explicitly facilitates interaction between vision and language features at shallow layers and supports modeling visual relationships relevant to the target actions. Comprehensive experiments on two datasets, JHMDB and UCF101-24, demonstrate the effectiveness of our framework, surpassing the state-of-the-art performance of the existing ZSSTAD method.

References

1. Chen, S., et al.: Watch only once: an end-to-end video action detection framework. In: ICCV (2021)
2. Cherti, M., et al.: Reproducible scaling laws for contrastive language-image learning. In: CVPR, pp. 2818–2829 (2023)
3. Faure, G.J., Chen, M.H., Lai, S.H.: Holistic interaction transformer network for action detection. In: WACV, pp. 3340–3350 (2023)
4. Ge, Z., Liu, S., Wang, F., Li, Z., Sun, J.: Yolox: exceeding yolo series in 2021. ArXiv arxiv:2107.08430 (2021)
5. Huang, W., Yeh, J.H., Faure, G.J., Chen, M.H., Lai, S.H.: Interaction-aware prompting for zero-shot spatio-temporal action detection. In: ICCVW, pp. 284–293 (2023)
6. Huang, X., Zhou, H., Yao, K., Han, K.: Froster: frozen clip is a strong teacher for open-vocabulary action recognition. ArXiv arxiv:2402.03241 (2024)
7. Jhuang, H., Gall, J., Zuffi, S., Schmid, C., Black, M.J.: Towards understanding action recognition. In: ICCV, pp. 3192–3199 (2013)
8. Jia, C., et al.: Scaling up visual and vision-language representation learning with noisy text supervision. ArXiv arxiv:2102.05918 (2021)
9. Li, L.H., et al.: Grounded language-image pre-training. In: CVPR, pp. 10955–10965 (2022)
10. Lin, T.Y., Dollár, P., Girshick, R.B., He, K., Hariharan, B., Belongie, S.J.: Feature pyramid networks for object detection. In: CVPR, pp. 936–944 (2017)
11. Lin, T.-Y., et al.: Microsoft COCO: common objects in context. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) ECCV 2014. LNCS, vol. 8693, pp. 740–755. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-10602-1_48
12. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning. In: NeurIPS, vol. 36 (2024)
13. Liu, S., et al.: Grounding dino: marrying dino with grounded pre-training for open-set object detection. ArXiv arxiv:2303.05499 (2023)
14. Ni, B., et al.: Expanding language-image pretrained models for general video recognition. In: ECCV 2022 (2022)
15. Pan, J., Chen, S., Shou, M.Z., Liu, Y., Shao, J., Li, H.: Actor-context-actor relation network for spatio-temporal action localization. In: CVPR (2021)
16. Radford, A., et al.: Learning transferable visual models from natural language supervision. In: ICML (2021)
17. Ren, S., He, K., Girshick, R.B., Sun, J.: Faster r-cnn: towards real-time object detection with region proposal networks. IEEE TPAMI **39**, 1137–1149 (2015)
18. Soomro, K., Zamir, A., Shah, M.: Ucf101: a dataset of 101 human actions classes from videos in the wild. ArXiv arxiv:1212.0402 (2012)
19. Sun, K., Xiao, B., Liu, D., Wang, J.: Deep high-resolution representation learning for human pose estimation. In: CVPR (2019)
20. Tang, J., Xia, J., Mu, X., Pang, B., Lu, C.: Asynchronous interaction aggregation for action detection. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J.-M. (eds.) ECCV 2020. LNCS, vol. 12360, pp. 71–87. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-58555-6_5

21. Vaswani, A., et al.: Attention is all you need. In: NeurIPS, vol. 30 (2017)
22. Wang, J., et al.: Deep high-resolution representation learning for visual recognition. IEEE TPAMI **43**, 3349–3364 (2020)
23. Wang, M., Xing, J., Mei, J., Liu, Y., Jiang, Y.: Actionclip: adapting language-image pretrained models for video action recognition. IEEE TNNLS (2023)
24. Wu, T., Cao, M., Gao, Z., Wu, G., Wang, L.: Stmixer: a one-stage sparse action detector. In: CVPR (2023)
25. Xie, S., Girshick, R.B., Dollár, P., Tu, Z., He, K.: Aggregated residual transformations for deep neural networks. In: CVPR, pp. 5987–5995 (2017)
26. Xue, H., et al.: Clip-vip: adapting pre-trained image-text model to video-language alignment. In: ICLR (2022)
27. Zhang, Y., et al.: Bytetrack: multi-object tracking by associating every detection box. In: ECCV (2022)
28. Zhao, J., et al.: Tuber: tubelet transformer for video action detection. In: CVPR, pp. 13598–13607 (2022)

Nonlinear Progressive Denoising: A Universal Regularized Denoising Strategy for Low PSNR Images

Ziyu Wu[1,2](✉) and Silong Peng[1,2]

[1] Institute of Automation, Chinese Academy of Sciences, Beijing, China
wuziyu2017@ia.ac.cn

[2] The University of Chinese Academy of Sciences, Beijing, China

Abstract. Traditional progressive strategy for denoising cascades a series of backbone denoisers to enhance the performance. However for denoising task with low peak signal-to-noise ratio(PSNR), we find this strategy is ineffective. Thus we extend the traditional progressive strategy to nonlinear progressive strategy learning to dig non-noise component from discarded noise of backbone denoisers. Inspired by the workflow of archaeology, the proposed strategy alternatively and repeatedly implement backbone denoiser and non-noise component digging module in a progressive manner. For traditional and deep denoisers, experiments show that for low PSNR images with regular shapes, the proposed strategy is able to help backbone denoisers recover these shapes with better discriminability than traditional progressive strategy. Although experiments find the proposed strategy help them achieve better performance on several public datasets with clear-cut rules, we make no claim that these published methods accompanied by our strategy will beat the state-of-the-art current algorithms on these and other natural image datasets. The novelty is that the proposed strategy is general and interpretable which can be applied to various deep or traditional denoisers for stronger nonlinear fitting capability and reliable performance improvement on severely ill-posed low PSNR noise removal problem.

Keywords: Nonlinear progressive strategy · Low PSNR denoising · Non-noise component digging

1 Introduction

Images with low PSNR are commonly encountered in many areas. These images may contain various sources of noise caused by hardware imperfection, complex observation environment and high-loss information channel, while the target information source may be weak compared with the noise. For this problem various proposed traditional or deep methods may fail to remove pure noise without mixing non-noise components. Although traditional progressive denoising strategy may alleviate this problem, we generalize the traditional strategy to its nonlinear version and achieve much better performance on several proof-of-principle experiments.

A. Antonacopoulos et al. (Eds.): ICPR 2024, LNCS 15329, pp. 245–259, 2025.
https://doi.org/10.1007/978-3-031-78110-0_16

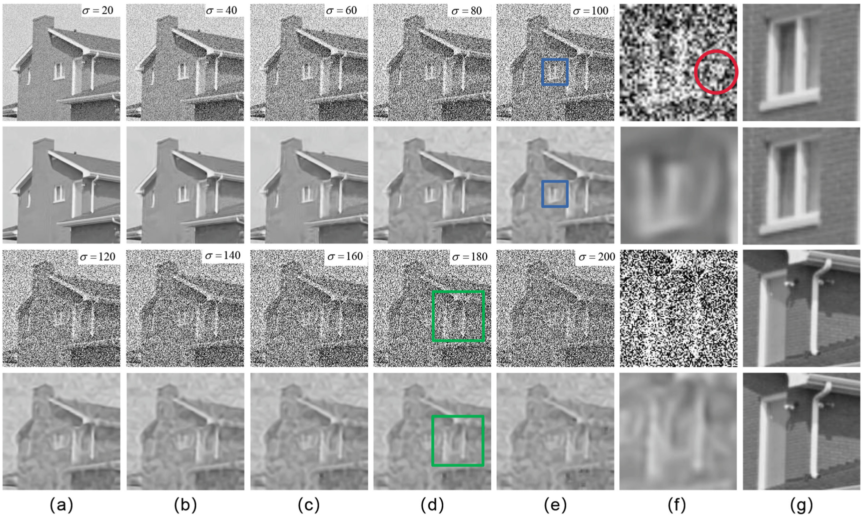


Fig. 1. Denoising performance of BM3D on different intensity of noise. The first and third rows are noisy images while the second and last rows are corresponding denoised images. The last two columns are the enlarged view in the colored box and the clean image respectively.

While a clean image is contaminated by strong noise, state-of-the-art denoisers may mistake the smooth component of noise as the component of clean image and discard the oscillating component of the clean image as noise. This effect will generate many denoised patches which do not satisfy the prior for the clean image. As shown in Fig. 1, BM3D effectively remove noise in high and medium PSNR noisy images. However when the PSNR decreases, the denoising performance drops sharply. As shown in Fig. 1(f) and Fig. 1(g), the fluctuation of noise misleads BM3D to generate nonexistent structures compared with the groundtruth clean image. The denoised image looks more similar to the noisy image rather than the groundtruth because the non-zero piecewise smooth part of the noise patch is mixed into the denoised clean image.

The main drawback of these denoisers is that strong noise will deteriorate their discrimination ability between noise and clean image. As a conceptual analogy to low PSNR image denoising task, archaeologists discriminate carefully between the relics and soil using small brushes progressively. After clearing most of the soil and observing the profile of the relics, they obtain more information and then adjust their strategy to dig out the relics more carefully. Inspired by this strategy, we study the possible method transfer from archaeology to low PSNR image denoising tasks. Although progressive denoising is not a brand new strategy for image processing, former traditional methods simply add part of discarded noise (via multiplying a scalar factor) back to the denoised image of backbone denoisers and repeatedly perform the backbone denoisers. Former deep learning based methods cascade various denoising blocks and train the whole network in an end-to-end way, which can be viewed as a unwrapping version of traditional progressive strategy. In this paper, we design corresponding nonlinear progressive denoising strategy for both traditional and deep backbone denoisers.

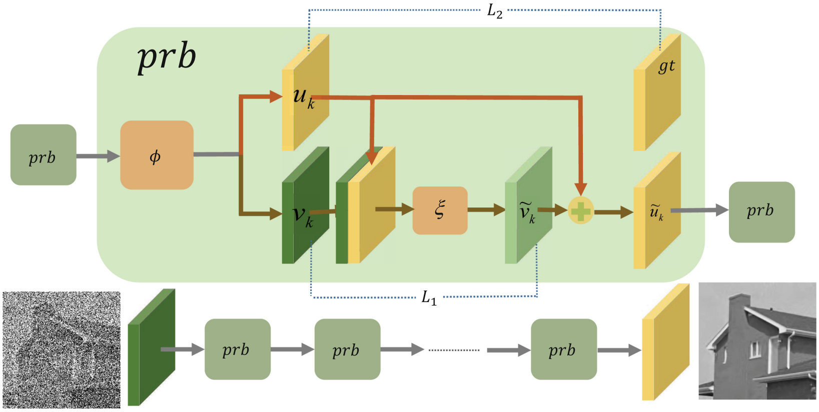


Fig. 2. Schematic diagram of our strategy. Progressive denoising block is abbreviated as *prb*, ϕ is the backbone denoiser, ξ is the non-noise component digging module. For deep backbone denoisers, *gt* is the given groundtruth for training, which is not necessarily clean image because Noise2Noise uses noisy image as groundtruth while Noise2Self uses the input as groundtruth. Since the strategy is general for traditional and deep backbone denoisers, the cuboid are images rather than convolution for better visualization. u_k, v_k, $\tilde{v}_k$, $\tilde{u}_k$ are denoised clean image, discarded noise, weakened noise to be added back and the noisy image with weakened noise.

2 Related Work

Single-stage image denoising refers to denoising a noisy image with a single denoiser and only one time. Progressive denoising strategy refers to denoising a noisy image with multiple denoisers and many times. With regard to single-stage image denoising, early research can be classified into four categories: variational functional based methods [4,14], sparse coding and low rank based methods [1,6,12,13,20], Non-local methods [3,5] and learning-based methods [2,10,11,17,19]

Knaus et al. [9] first propose progressive image denoising by reducing noise with deterministic annealing. As an improvement, Thote et al. [16] propose to estimate the noise variance self-adaptively and progressively remove the noise following the strategy of deterministic annealing. As an effective strategy, progressive strategy is also used in multi-stage image restoration [18] and other areas. However, these methods are not suitable for recovering images with low PSNR.

3 Proposed Method

The schematic diagram of the proposed strategy is shown in Fig. 2. Our strategy is a chain-like structure made up with several cascaded progressive denoising blocks, which is abbreviated as *prb*. One progressive denoising block contains a backbone denoiser ϕ and a non-noise component digging module ξ. We denote the denoised clean image and discarded noise of backbone denoiser ϕ at stage k as u_k and v_k respectively. The output of non-noise component digging module $\xi(u_k, v_k)$ is denoted as $\tilde{v}_k$(noise to be added back) and we denote $u_k + \tilde{v}_k$ as $\tilde{u}_k$(noisy image with weakened noise). The groundtruth clean image is u_{gt} while the groundtruth pure noise is v_{gt}. We define $\Delta(u_k, v_k) = \alpha v_k - \tilde{v}_k$ referring to the digged non-noise component.

3.1 Overall Structure

Backbone denoiser ϕ: For traditional backbone denoisers, we choose BM3D. For deep backbone denoisers, we consider two categories: trained without groundtruth clean image (Noise2Self [2] and Noise2Noise [11]) and with groundtruth clean image(Noise2Clean). The structures of Noise2Clean, Noise2Self and Noise2Noise are the same while their loss funtions are different.

Non-noise component digging module ξ: For traditional backbone denoisers, we adopt $(-1, 2)$ norm and Expected Patch Log Likelihood (EPLL) as the non-noise component digging module. For deep backbone denoisers, We adopt Unet [7,15] with residual connections to learn the non-noise component digging from v_k to $\tilde{v}_k$. We assign 0.9 to the energy decay factor α and regulate the energy decay between adjacent stages in the proposed strategy. We define $\tilde{u}_1 = u_1 + v_1, \tilde{u}_2 = (1-\alpha)\phi(\tilde{u}_1) + \alpha\tilde{u}_1 - \Delta(u_1, v_1), \cdots, \tilde{u}_n = (1-\alpha)\phi(\tilde{u}_{n-1}) + \alpha\tilde{u}_{n-1} - \Delta(u_{n-1}, v_{n-1}), n \in [1, N]$. Then we aim to find $\alpha \in [0, 1]$, $s \in [1, N]$ and function Δ such that for specific denoiser ϕ: $var_{u_s}(\tilde{\theta}) \leq var_{u_1}(\tilde{\theta}), \mathbb{E}_{u_s}(\tilde{\theta}) = \mathbb{E}_{u_1}(\tilde{\theta})$.

θ is the groundtruth clean image and $\widetilde{\theta}(u)$ be an unbiased estimator of θ. Figure 3(left) illustrates the denoising process of traditional progressive denoising strategy and the description is articulated in the caption of Fig. 3. Traditional progressive denoising strategy simply assumes $\Delta = 0$ which is a special case of the proposed nonlinear strategy. In Fig. 3(right), our strategy adds non-noise component digging module (dotted circle and array) into the traditional strategy. The proposed strategy can learn Δ from datasets or single image and model the prior of the data for better denoising performance.

3.2 Loss Function for Deep Backbone Denoisers

The loss function contains two parts, L_1 regulates the energy decay between v_k and $\tilde{v}_k$ while L_2 measures the distance between u_k and groundtruth. The *mask* is all-one matrix for Noise2Noise and Noise2Clean while blind-spot matrix for Noise2Self. The loss functions for these denoisers are as follows:

$$L = L_1 + \lambda L_2 \tag{1}$$

$$L = \sum_{k=2}^{s} ||\tilde{v}_k - \alpha v_k||_2^2 + \sum_{k=1}^{s} \lambda_k ||(u_k - GT) \times mask||_2^2 \tag{2}$$

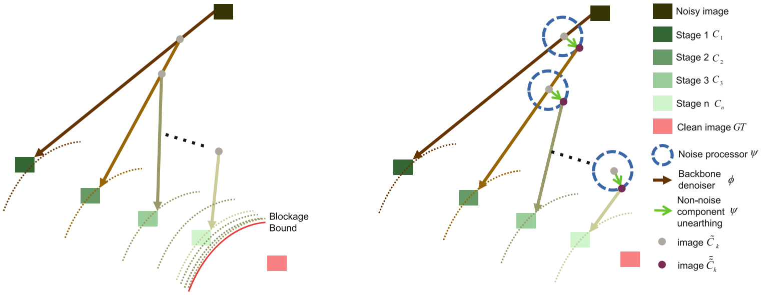


Fig. 3. The denoising process of ideal traditional progressive denoising strategy(left) and our progressive denoising strategy(right). The array represents backbone denoiser ϕ and the gray point represents the linear interpolation of noisy image and its denoised image. The circle and green array represent our non-noise component digging module. The dotted curves represents the relative distance between denoised images and the groundtruth clean image. The rectangles represent images (noisy images, denoised images, interpolated images and the groundtruth clean image). (Color figure online)

3.3 Algorithm for Traditional Backbone Denoisers

Although EPLL provides good priors for denoising a noisy image, balancing $||\delta||_2^2$ and $p(u - \Delta)$ via choosing suitable weight λ and optimizing $\lambda||\Delta||_2^2 + p(u - \Delta)$ is a hard problem. The reasons are twofold: (1) Only optimizing for u without considering v will generate $\tilde{v}$ mixing many non-noise components. (2) The weight λ is difficult to estimate because different choice of λ will significantly influence the optimization. In order to solve these problems, we propose to consider u and v simultaneously. We denote $v + \Delta$ as δ and consider the following optimization problem as the non-noise component digging module:

$$\min_{\delta} \lambda||\delta||_{-1,2} - \ln(p(z^i)) + \sum_i \frac{\beta}{2}||P_i(u + v - \delta) - z^i||^2 \tag{3}$$

p is posterior probability while P is sampling operator. According to the definition of $(-1, 2)$ norm and Parseval identity, $||\delta||_{-1,2}$ can be computed via the following equation:

$$||\delta||_{-1,2} = \sqrt{(-\delta, \Delta^{-1}\delta)} = \sqrt{(-\mathcal{F}(\delta), \mathcal{F}(\Delta^{-1}\delta))} \tag{4}$$

$$\mathcal{F}(\delta)(p,q) = \sum_{m=0}^{N-1}\sum_{n=0}^{N-1} \delta(m,n) e^{-j(2\pi/N)pm} e^{-j(2\pi/N)qn} \tag{5}$$

$$\mathcal{F}(\varDelta^{-1}\delta)(p,q) = \frac{1}{2(\cos(\frac{2p\pi}{N}) + \cos(\frac{2q\pi}{N}) - 2)} \mathcal{F}(\delta)(p,q) \tag{6}$$

$$(-\mathcal{F}(\delta), \mathcal{F}(\varDelta^{-1}\delta)) = \sum_{p=0}^{N-1}\sum_{q=0}^{N-1} \frac{\mathcal{F}(\delta)(p,q)^2}{2(2 - \cos(\frac{2p\pi}{N}) - \cos(\frac{2q\pi}{N}))} \tag{7}$$

$$\frac{\partial ||\delta||_{-1,2}}{\delta(m,n)} = \frac{\sum\limits_{p=0}^{N-1}\sum\limits_{q=0}^{N-1} \frac{\mathcal{F}(\delta)(p,q) e^{-j(2\pi/N)pm} e^{-j(2\pi/N)qn}}{(2-\cos(\frac{2p\pi}{N}) - \cos(\frac{2q\pi}{N}))}}{2\sqrt{(-\mathcal{F}(\delta), \mathcal{F}(\varDelta^{-1}\delta))}} \tag{8}$$

$$-\beta \sum_i (P_i^T P_i(u + v - \delta) - P_i^T z^i) + \lambda \frac{\partial ||\delta||_{-1,2}}{\delta(m,n)} = 0 \tag{9}$$

We denote $\mathcal{F}(\delta)(p,q)/(2 - \cos(\frac{2p\pi}{N}) - \cos(\frac{2q\pi}{N})$ as $\tilde{\delta}(p,q)$ and then Eq. (8) can be converted to the following equation:

$$\frac{\partial ||\delta||_{-1,2}}{\delta(m,n)} = \frac{\mathcal{F}(\tilde{\delta})}{2\sqrt{(-\mathcal{F}(\delta), \mathcal{F}(\varDelta^{-1}\delta))}} \tag{10}$$

Since it is difficult to obtain the closed-form solution for Eq. (9), we denote the δ of the previous step as $\hat{\delta}$ and approximate the solution of Eq. (9) with Eq. (11). As a typical fixed-point equation, we repeat Eq. (11) several times via assigning the latest δ to $\hat{\delta}$ and computing a new δ with given $\hat{\delta}$. In order to optimize Eq. (3), we alternatively solve for z^i given δ and solve for δ given z^i for several iterations. Then we increase β and continue to the next iteration.

$$y = (\beta \sum_i P_i^T P_i(u + v) - \beta P_i^T z^i - \frac{\lambda \partial ||\hat{\delta}||_{-1,2}}{\hat{\delta}(m,n)})$$

$$\delta = (\beta \sum_i P_i^T P_i)^{-1} y \tag{11}$$

In order to optimize z^i, we use GMM model in the log likelihood $\log(p(z^i))$ as shown in Eq. (12). We denote the right part in Eq. (13) as $g^t(\tilde{u}, z, \beta)$ where $\tilde{u} = u + v - \delta$. We use EM algorithm to minimize $g^t(\tilde{u}, z, \beta)$ due to the difficulty of obtaining the closed-form solution for minimizing $g^t(\tilde{u}, z, \beta)$.

$$\ln(p(z^i)) = \ln \sum_{k=0}^{K} \pi_k N(z^i|\mu_k, \sigma_k) \tag{12}$$

$$\begin{aligned}\min g^t(\tilde{u}, z, \beta) = \min \sum_i (\frac{\beta}{2}||P_i(u + v - \delta) - z^i||^2 \\ - \sum_{k=0}^{K} \omega_{i,k}^t \ln \frac{\pi_k N(z^i|\mu_k, \sigma_k)}{\omega_{i,k}^t})\end{aligned} \tag{13}$$

$$\begin{aligned}0 = \frac{\partial g^t}{\partial z^i} = -\beta(P_i\tilde{u} - z^i) + \sum_{k=0}^{K} \omega_{i,k}^t \sigma_k^{-1}(z^i - \mu_k) \\ = (\sum_{k=0}^{K} \omega_{i,k}^t \sigma_k^{-1} + \beta)z^i - (\beta P_i\tilde{u} + \sum_{k=0}^{K} \omega_{i,k}^t \sigma_k^{-1}\mu_k)\end{aligned} \tag{14}$$

$$\omega_{i,k}^t = \frac{\pi_k N(z^i|\mu_k, \sigma_k)}{\sum\limits_{k=0}^{K} \pi_k N(z^i|\mu_k, \sigma_k)} \tag{15}$$

In this paper we set $\lambda = 10$ and $\beta = 10$. Hard-thresholding for ω will degenerate the optimization to Wiener filtering combined with dual norm minimizing. Whether to use hard-thresholding varies for different images and for regular shapes we use hard-thresholding in this paper.

4 Experiments

4.1 Implementation Details

Dataset. For traditional backbone denoisers, we train the GMM on generated SEM cross(see Fig. 6) via ARTIMAGEN provided by National Institute of Standards and Technology [8] and train the GMM model with 100 mixtures and patch size 8. Then we create nine different crosses with random fine structures, edge effect and size for performance evaluation. For deep leanring based backbone denoisers, three public datasets with regular shapes are used. In detail, we use the same dataset Hanzi and MNIST with Noise2Self. We only consider addictive white Gaussian noise(AWGN) with varied noise levels $\sigma \in [0.5, 1, 2, 3]$ for Hanzi, $[0.2, 0.4, 0.6, 0.8]$ for MNIST dataset. In addition to the above two character datasets, we also consider an industrial dataset Gold-on-Carbon(GoC) dataset created randomly via the software ARTIMAGEN including 38227 images with size 64×64, which models the realistic physical effects of Scanning Electron Microscope including drifting, blurring, vibration, edge effect and noise process.

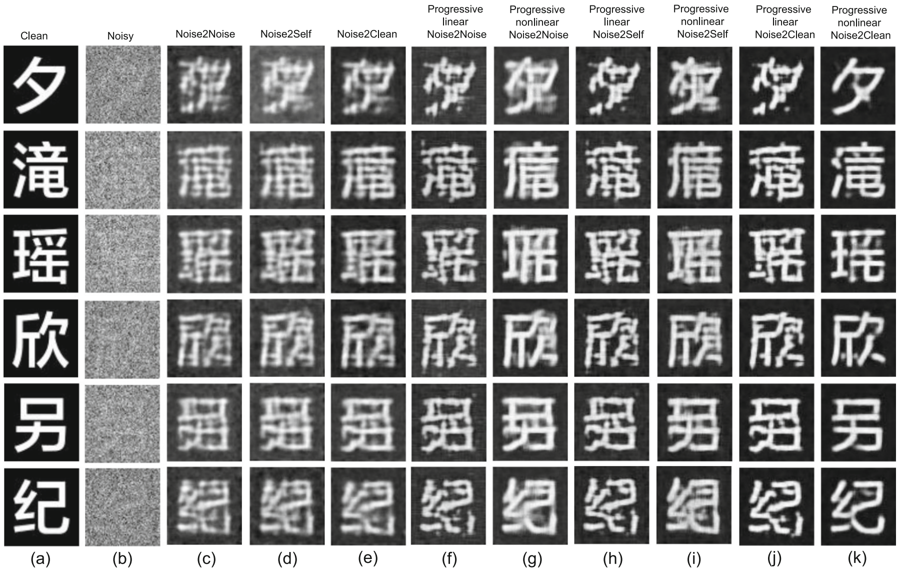

Clean
Noisy
Noise2Noise
Noise2Self
Noise2Clean
Progressive linear Noise2Noise
Progressive nonlinear Noise2Noise
Progressive linear Noise2Self
Progressive nonlinear Noise2Self
Progressive linear Noise2Clean
Progressive nonlinear Noise2Clean
(a)
(b)
(c)
(d)
(e)
(f)
(g)
(h)
(i)
(j)
(k)

Fig. 4. Column (a) are clean hanzi images and column (b) are corresponding noisy images with noisy variance $\sigma = 2$ (i.e. PSNR = -6.02 dB). (c), (d) and (e) are the denoising results of the backbone denoiser Noise2Noise, Noise2Self and Noise2Clean respectively. (f), (h) and (j) are the results of traditional progressive denoising strategy, termed as progressive linear Noise2Noise, Noise2Self and Noise2Clean respectively. (g), (i) and (k) are the results of our proposed deep progressive denoising framework, termed as progressive nonlinear Noise2Noise, Noise2Self and Noise2Clean respectively.

The clean images are generated using the default parameters of ARTIMAGEN and the following noise is added:

$$(Q_G + Q_P\sqrt{C_2})Gaussian(0,1)$$

where C_2 is clean image while Q_G and Q_P are chosen as $[0.1, 0.2, 0.3, 0.4]$ in the same. Since GoC dataset is not standard, the experiments results of this created dataset can be found in the additional material as a demo for realistic application and an further example for the proposed strategy.

Algorithm Details. We use the official implementation of BM3D, Noise2Self, Noise2Noise and Noise2Clean as our backbone denoisers. The backbone denoisers without progressive strategy are tested and the results are compared with the results obtained using traditional and our proposed progressive denoising strategy. We empirically use stage 9 for Hanzi dataset, 4 for MNIST dataset and GoC dataset. We randomly pick out 80 percent of Hanzi, MNIST and GoC datasets as training set, while the remainder are prepared for the test set. We use 20 stage progressive denoising(decay $\alpha = 0.95$) with BM3D and the proposed $(-1, 2)$ norm/EPLL as our backbone denoiser and non-noise component digging module.

The batch size for training is set to 32, 32 and 128 for Hanzi, GoC and MNIST datasets respectively while Adam optimizer is used with an self-adaptive learning rate with initial 10^{-3} for the deep nonlinear progressive denoising. We train the deep nonlinear progressive denoising strategy with 30 epochs. The hyper-parameter λ_k used to control the relative significance of several terms in the loss function is set to k which increases with the progressive stage k in the experiments.

Table 1. PSNR and SSIM values for Hanzi, MNIST and GoC datasets with backbone, 9 stage linear and 9 stage nonlinear strategy.

Denoising mode	Noise2Clean			Noise2Self			Noise2Noise		
	backbone	linear	nonlinear	backbone	linear	nonlinear	backbone	linear	nonlinear
Hanzi PSNR	11.74	11.42	**12.69**	11.52	11.26	**11.97**	11.4	11.26	**11.68**
Hanzi SSIM	0.43	0.52	**0.61**	0.37	0.42	**0.47**	0.37	0.43	**0.49**
MNIST PSNR	14.19	14.25	**16.02**	13.90	11.50	**15.20**	14.66	14.85	**16.18**
MNIST SSIM	0.45	0.51	**0.64**	0.41	0.19	**0.52**	0.49	0.56	**0.62**
GoC PSNR	26.68	19.96	**26.71**	23.06	18.56	**25.10**	**26.75**	20.71	24.38
GoC SSIM	0.76	0.82	**0.87**	0.43	0.31	**0.69**	**0.80**	0.35	0.45

The intermediate results for Hanzi dataset are shown in Fig. 7. For realistic application such as GoC, training data can be collected and nonlinear progressive denoising can be applied to enhance the performance. We generate the dataset and separate the training set(80%) and test set(20%) randomly. Experiment show the superiority except Noise2Noise scenario for GoC dataset. The reason may lie in that the backbone Noise2Noise is not robust for low PSNR images. Denoising performance is shown in Fig. 8.

4.2 Results for Deep Nonlinear Progressive Denoising Strategy

Results for Hanzi Dataset. As shown in Fig. 4(c)-(e), we find that Noise2Noise, Noise2Self and Noise2Clean fail to obtain satisfying denoising performance in the scenario of low PSNR situations. The PSNR values of traditional progressive denoising strategy are lower than backbone denoisers and we suppose the reason lies in that simply cascading backbone denoisers will generate nonexistent structures in the denoised images due to mistaking noise components

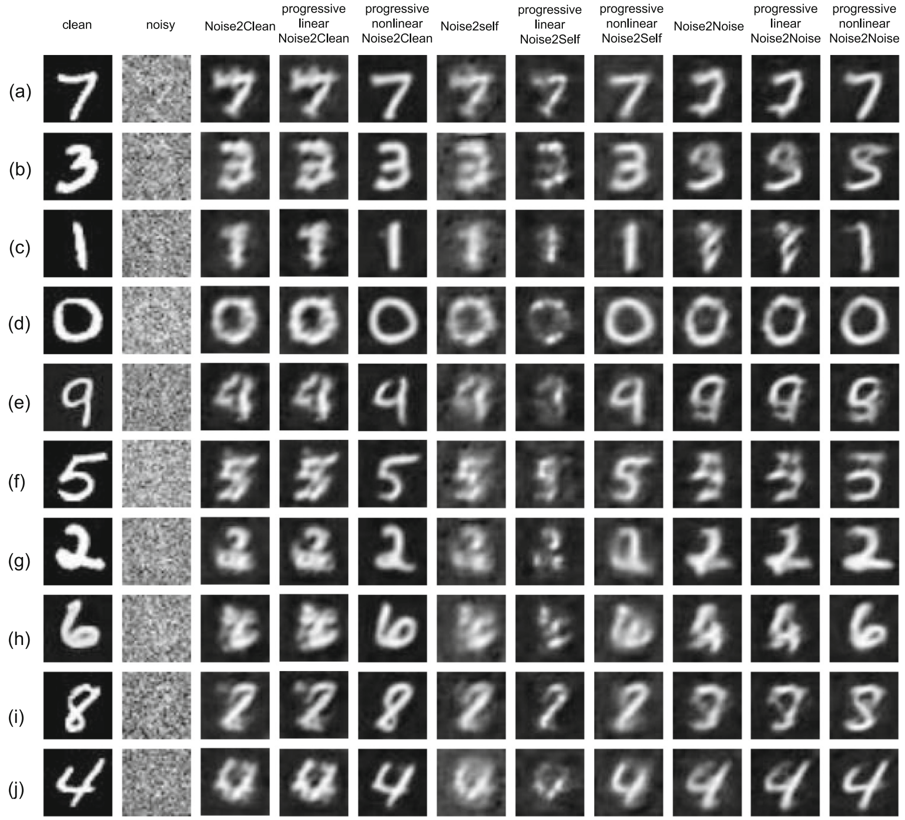

clean
noisy
Noise2Clean
progressive linear Noise2Clean
progressive nonlinear Noise2Clean
Noise2self
progressive linear Noise2Self
progressive nonlinear Noise2Self
Noise2Noise
progressive linear Noise2Noise
progressive nonlinear Noise2Noise
(a)
(b)
(c)
(d)
(e)
(f)
(g)
(h)
(i)
(j)

Fig. 5. Experiment results on MNIST. The first coloum is clean images and the second column is the corresponding noisy images with noise variance $\sigma = 0.6$ (i.e. PSNR $= -1.58$ dB). Traditional progressive denoising strategy is termed as progressive linear Noise2Noise, Noise2Self and Noise2Clean. Our proposed progressive denoising framework is termed as progressive nonlinear Noise2Noise, Noise2Self and Noise2Clean respectively.

and clean image components. Comparing the results in (f) and (g), (h) and (i), (j) and (k) of Fig. 4, non-noise component digging can effectively improve the denoising performance of traditional progressive denoising strategy. However in the Noise2Noise and Noise2Self denoising mode, since it is difficult to learn the prior of the dataset from noise contaminated images, progressive denoising strategy combined with non-noise digging module will sometimes mistake part of the noisy image due to the limitation of the learned prior. As shown in the second row of Fig. 4(g) and Fig. 4(i), the left part of the denoised image is distinct from that of Fig. 4(a) and Fig. 4(k).pg Further experiment results are shown in Fig. 7

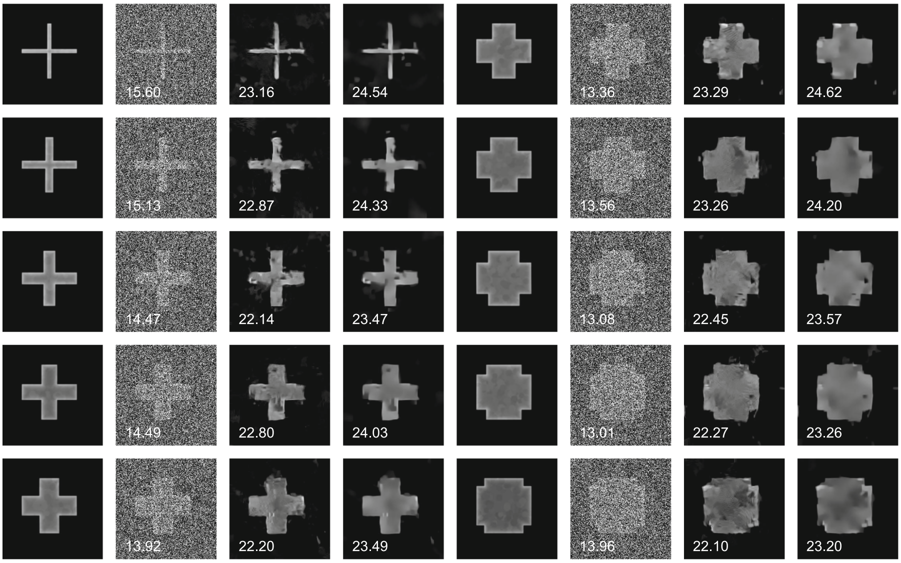
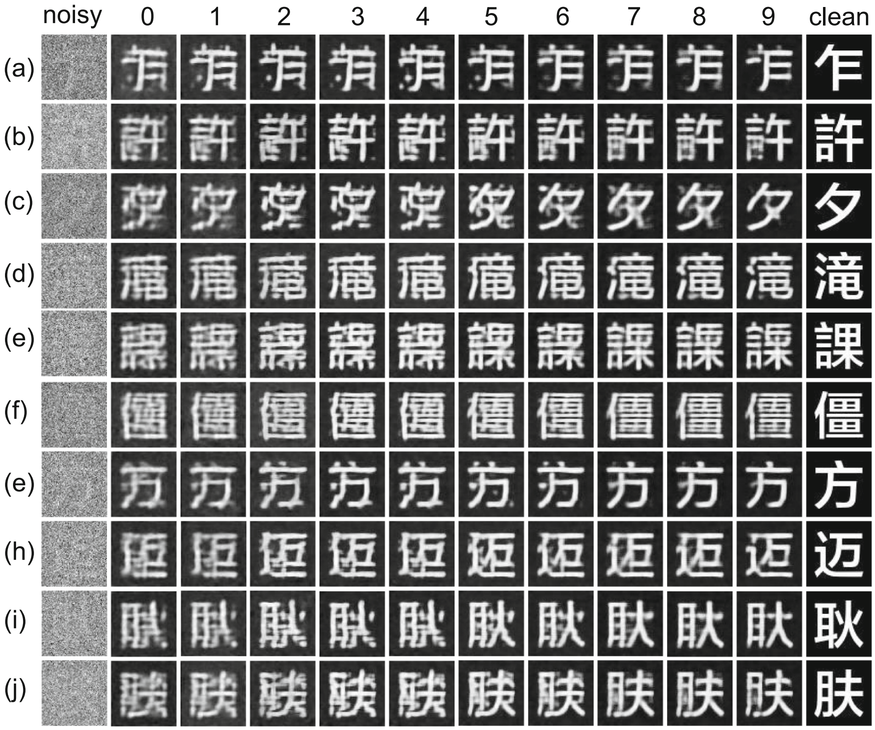


Fig. 6. Experiment results on SEM crosses. We use 20 stage progressive denoising with BM3D and the proposed $(-1, 2)$ norm/EPLL as our backbone denoiser and non-noise component digging module respectively. The first and fifth columns are the clean images while the second and fourth columns are noisy images with strong noise. The third and seventh columns are the denoising results of traditional progressive strategy with BM3D. The fourth and the last columns are the denoising results of the nonlinear progressive denoising strategy with BM3D. PSNR are marked in the images for comparisons.

of the additional material. We visualize the intermediate denoised images C_k in the trained 9-stage nonlinear progressive denoising strategy with Noise2Clean. Similar to the workflow of archaeologists to dig out cultural relics, the intermediate denoised images progressively approach the groundtruth clean image both from visual performance and PSNR/SSIM values.

Results for MNIST Dataset. Compared with Hanzi dataset, handwritten digits have more irregular structures. As shown in Fig. 5, our strategy achieves much better visual performance over backbone denoisers and traditional progressive denoising strategy. The performance improvements of traditional progressive denoising strategy over backbone denoisers in Noise2Clean and Noise2Noise training scenarios are quite limited. PSNR and SSIM are shown in Table 1 and our strategy achieves the highest scores on three training scenarios.



	noisy	0	1	2	3	4	5	6	7	8	9	clean
PSNR	-6.02	12.36	12.75	12.50	12.76	12.68	13.20	13.97	14.58	14.89	14.81	–
SSIM	0.037	0.42	0.44	0.49	0.56	0.55	0.60	0.62	0.63	0.65	0.68	1

Fig. 7. The intermediate denoised images in one model of trained 9-stage progressive denoising framework on Hanzi dataset. The first column are noisy images and the last column are groundtruth clean images. Column 2–11 are the intermediate denoised images u_k and PSNR/SSIM are listed below row (j). Since the noise component of $\widetilde{u}_k$ drops by 0.9 in one stage, the noise energy of $\widetilde{u}_9$ still remain 0.39 of that of $\widetilde{u}_1$. u_k is more suitable than $\widetilde{u}_k$ to demonstrate the conceptual analogy with archaeology.

4.3 Results for Nonlinear Progressive Denoising Strategy on Traditional Backbone Denoisers

As shown in Fig. 6, the first and fifth columns are the clean images while the second and sixth columns are noisy images with strong noise. The third and seventh columns are the denoising results of traditional progressive strategy with BM3D. The fourth and the last columns are the denoising results of the nonlinear progressive denoising strategy with BM3D. Experiments show that the proposed progressive strategy can evidently surpass the performance of traditional progressive strategy for various similar but different noisy samples.

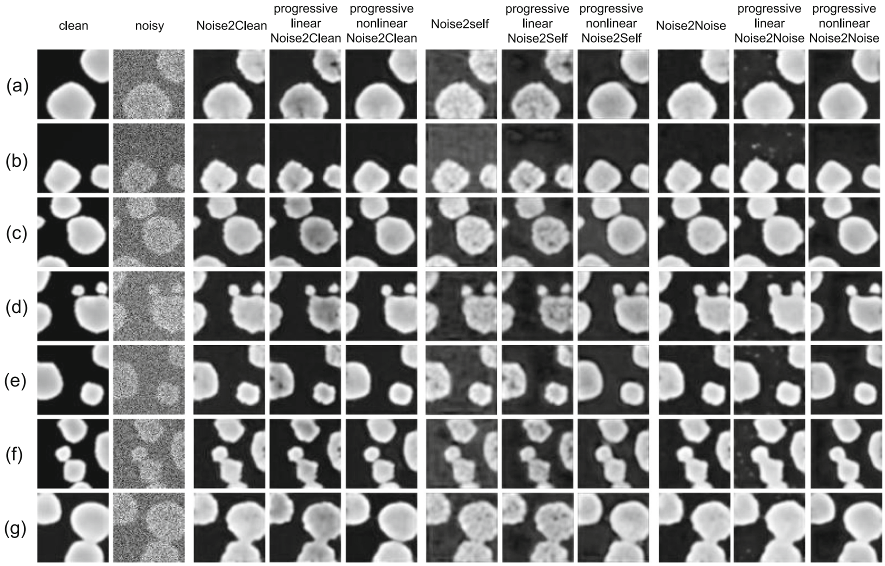

clean
noisy
Noise2Clean
progressive linear Noise2Clean
progressive nonlinear Noise2Clean
Noise2self
progressive linear Noise2Self
progressive nonlinear Noise2Self
Noise2Noise
progressive linear Noise2Noise
progressive nonlinear Noise2Noise
(a)
(b)
(c)
(d)
(e)
(f)
(g)

Fig. 8. Experiment results on GoC dataset. We use 4 stage progressive denoising with Noise2Clean, Noise2Self and Noise2Noise as our backbone denoisers. The second column is the corresponding noisy images with $Q_G = Q_P = 0.4$ (i.e. PSNR = 7.24 dB).

5 Conclusion

In this paper, we propose a nonlinear progressive denoising strategy without relying on specific backbone denoisers. Experiments show that our proposed strategy and non-noise component digging module can effectively enhance the denoising performance of backbone denoisers and traditional progressive denoising strategy. The main drawback of our framework is that the training dataset needs to share strong prior(similar features). In the future, we will develop better progressive strategy for more general denoising tasks.

Acknowledgments. The authors thank the Suzhou Research Center of the Institute of Automation, Chinese Academy of Science for funding this project.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Aharon, M., Elad, M., Bruckstein, A.M.: K-SVD: an algorithm for designing overcomplete dictionaries for sparse representation. IEEE Trans. Signal Process. **54**(11), 4311–4322 (2006)
2. Batson, J., Royer, L.: Noise2Self: blind denoising by self-supervision. In: Proceedings of the 36th International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 97, pp. 524–533. PMLR (2019)
3. Buades, A., Coll, B., Morel, J.: A non-local algorithm for image denoising. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 60–65 (2005)
4. Chambolle, A.: An algorithm for total variation minimization and applications. J. Math. Imaging Vis. **20**(1–2), 89–97 (2004)
5. Dabov, K., Foi, A., Katkovnik, V., Egiazarian, K.O.: Image denoising by sparse 3-D transform-domain collaborative filtering. IEEE Trans. Image Process. **16**(8), 2080–2095 (2007)
6. Gu, S., Zhang, L., Zuo, W., Feng, X.: Weighted nuclear norm minimization with application to image denoising. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 2862–2869 (2014)
7. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 770–778 (2015)
8. Čižmár, P., Vladár, A.E., Ming, B., Postek, M.T.: Simulated SEM images for resolution measurement. Scanning **305**, 381–91 (2008)
9. Knaus, C., Zwicker, M.: Progressive image denoising. IEEE Trans. Image Process. **23**, 3114–3125 (2014)
10. Krull, A., Buchholz, T., Jug, F.: Noise2Void - learning denoising from single noisy images. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 2129–2137 (2019)
11. Lehtinen, J., et al.: Noise2Noise: learning image restoration without clean data. In: Proceedings of the 35th International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 80, pp. 2971–2980. PMLR (2018)
12. Ma, J., Plonka-Hoch, G.: The curvelet transform. IEEE Signal Process. Mag. **27**, 118–133 (2010)
13. Mallat, S.: A theory for multiresolution signal decomposition: the wavelet representation. IEEE Trans. Pattern Anal. Mach. Intell. **11**, 674–693 (1989)
14. Parisotto, S., Lellmann, J., Masnou, S., Schönlieb, C.: Higher-order total directional variation: imaging applications. SIAM J. Imaging Sci. **13**(4), 2063–2104 (2020)
15. Ronneberger, O., Fischer, P., Brox, T.: U-Net: convolutional networks for biomedical image segmentation. In: Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F. (eds.) MICCAI 2015. LNCS, vol. 9351, pp. 234–241. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-24574-4_28
16. Thote, B.K., Jondhale, K.C.: Progressive image denoising using fast noise variance estimation. In: Proceedings of the Third International Symposium on Computer Vision and the Internet (2016)
17. Ulyanov, D., Vedaldi, A., Lempitsky, V.S.: Deep image prior. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 9446–9454 (2018)
18. Zamir, S.W., et al.: Multi-stage progressive image restoration. In: 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 14816–14826 (2021)

19. 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)
20. Zoran, D., Weiss, Y.: From learning models of natural image patches to whole image restoration. In: IEEE International Conference on Computer Vision, pp. 479–486 (2011)

Unified-EGformer: Exposure Guided Lightweight Transformer for Mixed-Exposure Image Enhancement

Eashan Adhikarla[1](✉), Kai Zhang[1], Rosaura G. VidalMata[2], Manjushree Aithal[2], Nikhil Ambha Madhusudhana[2], John Nicholson[2], Lichao Sun[1], and Brian D. Davison[1]

[1] Lehigh University, Bethlehem, PA 18015, USA
{eaa418,kaz321,lis221,bdd3}@lehigh.edu
[2] Lenovo Research, Morrisville, USA
{rosaurav,maithal,amnikhil,jnichol}@lenovo.com

Abstract. Despite recent strides made by AI in image processing, the issue of mixed exposure, pivotal in many real-world scenarios like surveillance and photography, remains a challenge. Traditional image enhancement techniques and current transformer models are limited with primary focus on either overexposure or underexposure. To bridge this gap, we introduce the Unified-Exposure Guided Transformer (**Unified-EGformer**). Our proposed solution is built upon advanced transformer architectures, equipped with local pixel-level refinement and global refinement blocks for color correction and image-wide adjustments. We employ a guided attention mechanism to precisely identify exposure-compromised regions, ensuring its adaptability across various real-world conditions. U-EGformer, with a lightweight design featuring a memory footprint (peak memory) of only ∼**1134 MB** (0.1 Million parameters) and an inference time of 95 ms (over **9x** faster than typical existing implementations, is a viable choice for real-time applications such as surveillance and autonomous navigation. Additionally, our model is highly generalizable, requiring minimal fine-tuning to handle multiple tasks and datasets with a single architecture.

Keywords: Computer Vision · Image Processing · Image Restoration · Low-Light Image Enhancement · Unified Learning

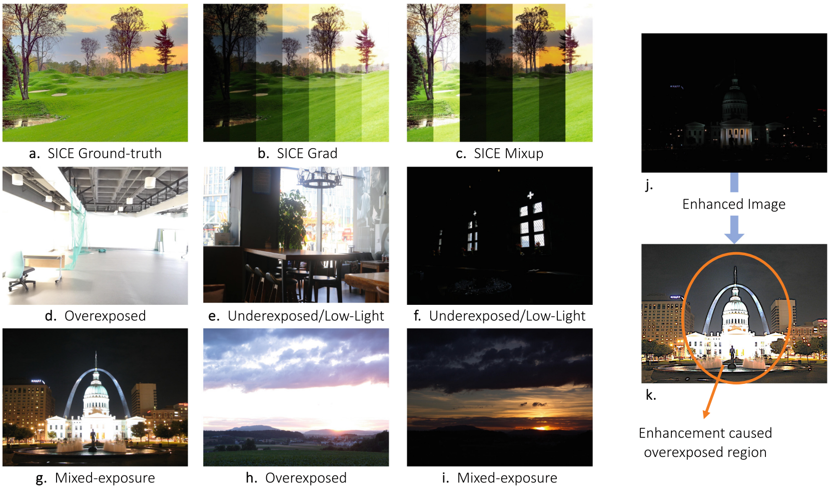
1 Introduction

AI-driven image processing has significantly broadened the scope for enhancing visual media quality. A critical challenge within low-light image enhancement is addressing ***mixed exposure*** in images as in Fig. 1 g., where a single frame contains both underlit[1] (below 5 lux, including underexposed) and overlit[2] (overexposed) regions. This

[1] insufficient brightness in an image where details are lost due to lack of signal.
[2] excessive brightness in an image where details are lost due to signal clipping or saturation.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/978-3-031-78110-0_17.

A. Antonacopoulos et al. (Eds.): ICPR 2024, LNCS 15329, pp. 260–275, 2025.
https://doi.org/10.1007/978-3-031-78110-0_17



a. SICE Ground-truth

b. SICE Grad

c. SICE Mixup

d. Overexposed

e. Underexposed/Low-Light

f. Underexposed/Low-Light

g. Mixed-exposure

h. Overexposed

i. Mixed-exposure

Fig. 1. Sub-figures [a,b,c] show the handcrafted mixed exposure dataset by Zheng et al. [74]; images [d-i] from Cai et al. [5] illustrate real-world scenarios of underexposure, overexposure, and mixed-exposures. Images [j,k] demonstrate the problem practically.

issue extends beyond academic interest and has significant real-world implications. For instance, in video calls (e.g., in cafeterias as in Fig. 1 e.) and live streaming, low-light enhancement is pivotal for clear visual communication. Other areas of application include autonomous driving, surveillance and security, photography, etc. Professional photographers often use high-end DSLR cameras and meticulously adjust settings such as aperture, ISO, and utilize specialized filters to mitigate exposure issues. However, such pre- and post-processing techniques are often not practical.

Existing methods for correcting mixed exposure images have typically treated underexposure [16,44,53] and overexposure as separate challenges within the low-light image enhancement task. Although these methods (such as FECNet [20], ELCNet+ERL [21], IAT [12]) have made progress, they commonly assume uniform scene illumination, leading to global adjustments that either brighten or darken the entire image. Such approaches falls short when dealing with images that have both overexposed and underexposed regions due to non-uniform lighting, resulting in suboptimal performance. For example, ZeroDCE [16] and RUAS [37] can worsen overexposure in background regions while trying to enhance underexposed foreground subjects as shown in Fig. 1 k.

We see this as a gap in the literature that has not been fully addressed, despite some efforts such as LCDNet [52] and night enhancement approaches [3,5,29]. The motivation for our work arises from the need to address these limitations with a solution that can handle mixed exposure scenarios effectively and is suitable for deployment on edge devices. An additional challenge when applying a low-light enhancement approach to a multi-exposed image is when a model trained solely on underexposed paired images inadvertently exacerbates overexposed regions, as seen in Fig. 1 k, and vice versa. Consequently, the issue of mixed exposure emerges as a pivotal yet largely unexplored frontier. Notably, mobile phone cameras face this issue acutely, requiring lightweight, lowlatency models that can operate within the device's resource constraints while delivering high-quality image enhancement. Our goal is to develop a solution that not only addresses these issues but also offers faster inference and a lower memory footprint, making it ideal for practical applications on edge devices.

This paper develops an effective and computing-efficient approach to tackle mixed exposure in low-light image enhancement. Our major contributions are as follows:

1. We introduce a novel *unified* framework within the transformer architecture that leverages **attention-** and **illuminance-maps** to enhance precision in processing affected regions at the pixel-level, addressing the challenges of underexposure, overexposure, and mixed exposure in images as a single task.
2. Our method achieves remarkable efficiency, with an average inference speed of **0.095 s** per image[3], significantly faster with lesser memory consumption than many existing frameworks. Coupled with a compact architecture of only **101 thousand** parameters, the model is ideal for deployment on edge devices.
3. We present a novel "**MUL-ADD**" loss function that intelligently combines contrast scaling and brightness shifting to adaptively enhance images, improving dynamic range and preventing over-smoothing.
4. We develop an Exposure-Aware Fusion (**EAF**) Block, designed for the efficient fusion of local and global features. This block refines image exposure corrections with heightened precision, enabling context-aware enhancements tailored to the specific exposure needs of each image region.

2 Related Work

Traditional to Advanced Deep Learning Techniques. In the realm of image enhancement and exposure correction, significant strides have been made to address the challenges posed by exposure scenarios. Early techniques [6,24,34] leveraged contrast-based histogram equalization (HE), laying the groundwork for more advanced methods. These initial approaches were followed by studies in Retinex theory, which focused on decomposing images into reflection and illumination maps [32,33]. The advent of deep learning transformed exposure correction, with a shift from enhancing low-light images to addressing both underexposure and overexposure [2,3,11,16,53,62,68,72]. The notable work of Afifi et al. [2] stands out, employing deep learning to simultaneously address underexposure and overexposure, a task not adequately considered by previous methodologies. There was a momentous shift towards convolutional neural network (CNN)-based methods, achieving state-of-the-art results and improving the accuracy and efficiency of exposure correction algorithms [16,30,44,50,53,63].

Addressing the Challenges of Mixed Exposure. Despite these advancements, the challenges of mixed exposure have remained relatively unaddressed in high-contrast scenarios. Benchmark datasets such as LOL [59], LOL-[4K;8K] [55], SID [8], SICE [5], and ELD [60] offer limited mixed exposure instances, highlighting a gap in both data and models. Synthetic datasets like SICE-Grad and SICE-Mix are enabling the

[3] computed over LOL-v2 test dataset following previous benchmarks [9,12,16,27].

development of methods tailored to mixed exposure scenarios [13,38,48,75]. Some representative methods of deep learning like RetinexNet [59] and KIND [72], focusing on illumination and reflectance component restoration in images, achieved good performance, but most methods emphasize either underexposure or overexposure correction, and fail to correct the combination. More recent studies have attempt to address the challenges of correcting both underexposed and overexposed images [52,64], a task complicated by the differing optimization processes required for each type of exposure. MSEC [3], a revolutionary work in this area, utilizes a Laplacian pyramid structure to incrementally restore brightness and details, dealing with a range of exposure levels.

To manage the correction of a wide range of exposures, several recent works were proposed, such as Huang et al. [18] using exposure normalization, CMEC [45] targeting exposure-invariant feature spaces for exposure consistency, and ECLNet [18] using bilateral activation for exposure consistency. Moreover, Wang et al. [52] tackle the issue of uneven illumination, while FECNet [19] uses a Fourier-based approach for lightness and structure. Still challenges remain unsolved: (1) handling nonuniform illumination, (2) simultaneously addressing both overexposure and underexposure within the same frame, and, (3) ensuring that global adjustments do not adversely affect local regions. Our work addresses these challenges by introducing a unified framework that uses attention and illuminance maps to process mixed exposure regions more precisely.

Emerging Trends with Computational Challenges. Recent studies have begun addressing the dual challenge of correcting under and overexposed images through innovative architectures, including transformers [14], despite their computational intensity as noted in works like Vaswani et al. [51]. The realm of image enhancement has seen remarkable models with human-level enhancement capabilities, such as ExposureDiffusion [58], Diff-retinex [66], wavelet-based diffusion [26], PyDiff (Pyramid Diffusion) [76], Global structure aware diffusion [17], LLDiffusion [54], and Maximal diffusion values [31], demonstrating significant advancements. However, these resource-intensive models, alongside the emerging vision-language models (VLMs) in image restoration, present deployment challenges on edge devices, often requiring tens of seconds to process a single HD or FHD image [41,56]. In contrast, Unified-EGformer achieves an average inference speed of $\sim$200 milliseconds on HD images, 9.6$\times$ faster on average inference time among representative models in Table 2 in Ye et al. [65].

3 Methodology: Unified-EGformer

Unifed-EGformer achieves image enhancement through an Attention Map Generation mechanism that identifies exposure adjustment regions, a Local Enhancement Block for pixel-wise refinement, a Global Enhancement Block for color and contrast adjustment, and an Exposure Aware Fusion (EAF) block that fuses features from both enhancements for balanced exposure correction as shown in Fig. 2.

3.1 Guided Map Generation

Unified-EGformer introduces significant advancements in the attention mechanism and feed-forward network within its architecture to adeptly handle mixed exposure scenarios. These enhancements are encapsulated as follows:

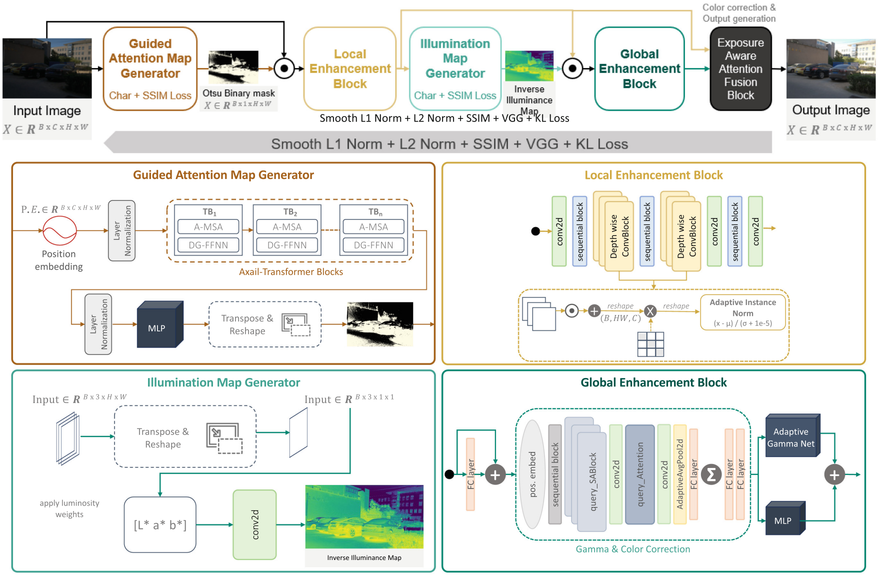
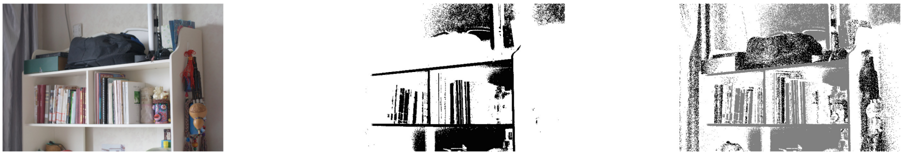


Fig. 2. U-EGformer's training, fine-tuning and inference pipelines. All four modules are showcased: Guided Attention Map, Local Block, Global Block, and Exposure Aware Fusion block.


(a) Input

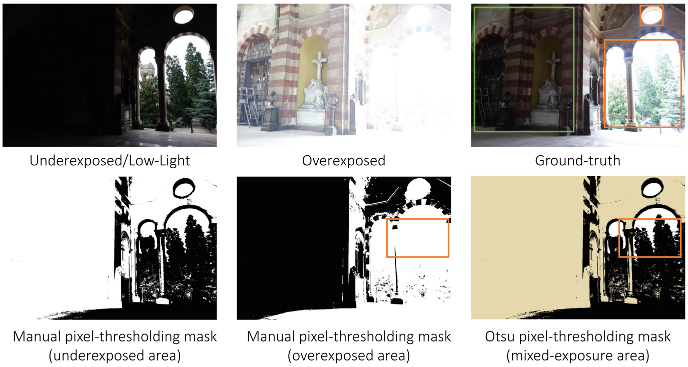
(b) Single Exposure

(c) Bi-Exposure

Fig. 3. Visualization of Otsu thresholding challenge: **(a)** original image, **(b)** mask for single exposure (underexposed), and **(c)** mask for bi-exposure (under and overexposed).

Thresholding. To highlight the sub-regions of the images with impacted exposure problems, we need a way to point out those impacted set of pixels within the input. We use Otsu thresholding [46], a traditional yet effective technique. It is a global thresholding technique for automatic thresholding that works by selecting the threshold to minimize *intra-class variance* (variance within a class) or maximize *inter-class variance* (variance between classes). However, this method induces granular noise in the image [46] due to the non-uniform pixels in low lux regions. The noise is highlighted by the resultant masks as shown in Fig. 3, and will influence the subsequent exposure correction.

To mitigate noise, we implemented adaptive thresholding using pixel blocks and downsampled images. We further reduced noise creep with nearest neighbor downsampling and Gaussian blur. Integrating Charbonnier loss [7] into our attention map mechanism encouraged smoother transitions in areas of high gradient variance, specifically targeting denoising. This component, combined with the SSIM loss that is applied directly on the input, synergistically contributes to noise reduction.


Underexposed/Low-Light

Overexposed

Ground-truth

Manual pixel-thresholding mask (underexposed area)

Manual pixel-thresholding mask (overexposed area)

Otsu pixel-thresholding mask (mixed-exposure area)

Fig. 4. The top row, from left to right: an underexposed image, an overexposed image, and the ground truth. The bottom row illustrates pixel-thresholding binary masks for the underexposed (white indicating underexposed regions), overexposed (white indicating overexposed regions) and Otsu thresholding for mixed exposures (yellow indicating underexposed regions, white representing overexposed areas, and black as correctly exposed portions. (Color figure online)

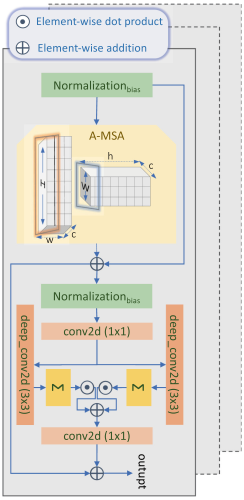


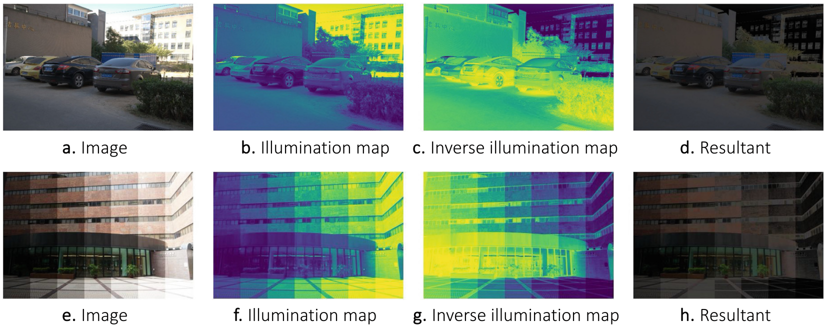
Fig. 5. Chain of efficient transformer blocks equipped with A-MSA, DGFN.

Our implementation of threshold selection is as follows. First, we calculate the Otsu average across the training set to establish a baseline for automatic thresholding. We then apply this average threshold to each image in the dataset. Using a data set-specific threshold, this method ensures a more uniform application[4] of the Otsu method.

Attention Map Generator. The Unified-EGformer begins with a guided attention map generator, designed to identify regions within an image affected by mixed exposure. This process involves generating a map $\boldsymbol{M_g} \in \mathbb{R}^{B\times H\times W\times C}$, where H, W, and C represent the height, width, and channel dimensions of the input image $x \in \mathbb{R}^{H\times W\times C}$. This map, $m \in \boldsymbol{M_g}$, is used in an element-wise dot product with the image, resulting in a guided input image that undergoes underexposed, overexposed, or mixed exposure enhancement, as depicted in Fig. 4, demonstrating how we apply Otsu thresholding to get attention masks labels.

The architecture incorporates improved Swin transformer [39] blocks, leveraging Axis-based Multi-head Self-Attention (A-MSA) and Dual Gated Feedforward Network (DGFN) [14,51,55] for efficient, fast and focused feature processing. The A-MSA reduces computational load by applying self-attention across height and width axes sequentially, optimizing for local contexts within high-resolution images, as illustrated in Fig. 5 (see details in Supplement Sect. 7.2 The DGFN introduces a dual gating mechanism to the feedforward network, allowing selective

[4] reducing noise propensity in low-light conditions and enhancing exposure correction consistency.



a. Image

b. Illumination map

c. Inverse illumination map

d. Resultant

e. Image

f. Illumination map

g. Inverse illumination map

h. Resultant

Fig. 6. Ground-truth **(a,e)** with it's corresponding illuminance map **[b,f]**, inverse illuminance maps **[c,g]**, and the enhanced results showcasing the effectiveness of the inverse illuminance normalization **[d,h]**. Note that the illuminance maps are using false color.

emphasis on critical features necessary to distinguish and correct underexposed and overexposed areas effectively (see Supplement Sect. 7.3) (Fig. 6).

Illumination Map Generator. We incorporate the generation of an illumination map I_{illum} into the global enhancement block, providing a foundational layer for exposure correction. Unlike the complex attention mechanisms required for a local block, generating an illumination map leverages a direct conversion from RGB to luminance ($I_{\text{rgb}} \rightarrow I_{\text{illum}}$), offering a simpler and faster solution ($I_{\text{illum}} = \mathbf{W} \times I_{\text{rgb}}$), where $\mathbf{W}$ corresponds to the luminosity method [49]. The block can utilize the illuminance information to dynamically adjust global parameters such as color balance and exposure levels, ensuring the enhancements are computationally efficient. Our ablation studies shown in Table 3 confirm that these enhancements are perceptually meaningful compared to the baseline without the illumination map.

3.2 Unified-Enhancement

Local Enhancement Block (LEB). The LEB takes a dot product of attention map $\mathcal{A}(x)$ and the sRGB $I(x)$ image, that uniquely forms an objective mapping from input to output. The LEB applies a lightweight convolutional block inspired from PE-Yolo [67], UWFormer [10] localized pixel-wise refinement. We utilize an adaptive instance-based color normalization for capturing a wider range of colors based on individual inputs. Unlike traditional methods, our approach maintains the integrity of features without down-sampling and up-sampling. The output of this block calculates multiplicative (M_L) and additive (A_L) correction factors through a feed-forward network, enabling precise pixel-wise enhancement. The local enhancement is formulated as:

$$\hat{I}(x) = M_L(x) \odot I(x) + A_L(x) \tag{1}$$

where $\hat{I}(x)$ is the locally enhanced image, $I(x)$ the original image, and $\odot$ denotes element-wise multiplication.

Global Enhancement Block (GEB). Complementary to the LEB, the global enhancement block adjusts the image's overall exposure through adaptive gamma correction and color balance. Unlike static bias adjustments [12], this block dynamically calculates the gamma correction factor and color transformation parameters based on the image's content. We implement a convolutional subnetwork with GELU activations and adaptive average pooling, followed by a sigmoid to automatically compute global parameters, denoted by θ. This adaptive approach to global enhancement allows for a more nuanced and content-aware balance of contrast and color. The global correction function is described as:

$$\mathrm{G}(I) = f_\gamma(I;\theta), \quad I \in \mathbb{R}^{H\times W\times C} \tag{2}$$

where $G(I)$ is the globally enhanced image, and f_γ represents the function for global adjustments, influenced by the calculated parameters θ.

3.3 Exposure-Aware Fusion (EAF) Block

Our novel exposure aware fusion block is architecturally designed to integrate local and global enhancement features, enabling comprehensive image enhancement. The fusion process begins with two convolutional layers that apply spatial filtering to extract the salient features necessary for exposure correction. We also use global average pooling, mapping the feature maps to the global context vector. These fusion weights serve as a gating mechanism to regulate the contribution of local and global features. They are adaptively learned, encapsulating both detailed texture information and broad illumination context.

3.4 Loss Functions

To enhance multiple aspects of image quality, our training uses a detailed loss function setup in the RGB color space. It includes L_1 and L_2 losses for handling outliers and preserving fine details, SSIM for maintaining structural integrity, and VGG for ensuring semantic consistency. We also incorporate a novel MUL-ADD loss to accurately adjust the image's contrast and brightness, ensuring that the dynamic range is well represented without blurring details. The VGG loss helps match the output to high-level visual quality standards.

MUL-ADD loss. The Multiplicative-Additive loss is specifically designed to handle mixed exposure scenarios by optimizing the multiplicative and additive adjustments applied during the enhancement process. This loss is mathematically defined as:

$$L_{\text{MA}}(M_L, A_L, I_{\text{low}}, I_{\text{high}}) = \boldsymbol{\xi}\ L_1(M_L, \hat{M}(I_{\text{low}}, I_{\text{high}})) + \boldsymbol{\psi}\ L_1(A_L, \hat{A}(I_{\text{low}}, I_{\text{high}})) \tag{3}$$

Where:

$$\hat{M}(I_{\text{low}}, I_{\text{high}}) = \frac{I_{\text{high}} - \hat{A}(I_{\text{low}}, I_{\text{high}})}{I_{\text{low}} + \epsilon}, \quad \hat{A}(I_{\text{low}}, I_{\text{high}}) = I_{\text{high}} - \frac{I_{\text{high}} \cdot \epsilon}{I_{\text{low}} + \epsilon} \tag{4}$$

In these equations, I_{low} and I_{high} represent the low-light input and the target high-quality image, respectively, while M_{L} and A_{L} are the multiplicative and additive components learned by the local enhancement block. The parameters $\boldsymbol{\xi}$ and $\boldsymbol{\psi}$ control the balance between the two components of the Mul-Add loss. The small constant $\epsilon = 1e - 8$ is added to avoid division by zero. This loss function ensures that the multiplicative and additive factors are optimized to produce a natural-looking enhancement without overcompensating in either direction. Our combined loss function L_{C}, considering both local and global outputs, is detailed below:

$$\begin{aligned}\mathcal{L}_{\text{C}}(y, \hat{y}, I_{low}, I_{high}) = \boldsymbol{\alpha}\, L_1(y, \hat{y})_{(l,g)} + \boldsymbol{\beta}\, L_2(y, \hat{y}) + \boldsymbol{\gamma}\, L_{\text{SSIM}}(y, \hat{y}) + \boldsymbol{\delta}\, L_{\text{VGG}}(y, \hat{y}) \\ + \boldsymbol{\eta}\, L_{\text{MA}}(M_L, A_L, I_{\text{low}}, I_{\text{high}}) + L_{\text{attn}}(M_g(b), M)\end{aligned} \tag{5}$$

where $\alpha, \beta, \gamma, \delta$, and η are hyperparameters balancing the influence of each loss term, y is the ground truth, $\hat{y}$ is the predicted image, M_L and A_L are the multiplicative and additive components of the local block, I_{low} is the low-light input, and I_{high} is the target high-quality image. Our fine-tuning stage's loss equation can be presented with the physics-based KL-divergence loss:

$$\mathcal{L}_{\text{finetune}}(y, \hat{y}, P, Q) = \boldsymbol{\lambda}\, L_1(y, \hat{y}) + \boldsymbol{\mu}\, L_{\text{SSIM}}(y, \hat{y}) + \boldsymbol{\nu}\, L_{\text{KL}}(\mathcal{P}, Q)_{(l,g)} \tag{6}$$

Table 1. Results for our exposure guided transformer approach over ME-v2 [3] and SICE-v2 [5] datasets. 🟥, 🟩, 🟦 denotes top three respectively. We did not include other recent models that are too complex ($> 10M$ params).

Method	ME-v2						SICE-v2						#params
	Underexposure		Overexposure		Average		Underexposure		Overexposure		Average		
	PSNR ↑	SSIM ↑	PSNR ↑	SSIM ↑	PSNR ↑	SSIM ↑	PSNR ↑	SSIM ↑	PSNR ↑	SSIM ↑	PSNR ↑	SSIM ↑	
RetinexNet [59]	12.13	0.6209	10.47	0.5953	11.14	0.6048	12.94	0.5171	12.87	0.5252	12.90	0.5212	0.840M
URetinexNet [61]	13.85	0.7371	9.81	0.6733	11.42	0.6988	12.39	0.5444	7.40	0.4543	12.40	0.5496	1.320M
Zero-DCE [16]	14.55	0.5887	10.40	0.5142	12.06	0.5441	16.92	0.6330	7.11	0.4292	12.02	0.5211	0.079M
Zero-DCE++ [35]	13.82	0.5887	9.74	0.5142	11.37	0.5583	11.93	0.4755	6.88	0.4088	9.41	0.4422	0.010M
DPED [25]	20.06	0.6826	13.14	0.5812	15.91	0.6219	16.83	0.6133	7.99	0.4300	12.41	0.5217	0.390M
KIND [72]	15.51	0.7115	11.66	0.7300	13.20	0.7200	15.03	0.6700	12.67	0.6700	13.85	0.6700	0.590M
DeepUPE [53]	19.10	0.7321	14.69	0.7011	16.25	0.7158	16.21	0.6807	11.98	0.5967	14.10	0.6387	7.790M
SID [8]	19.37	0.8103	18.83	0.8055	19.04	0.8074	19.51	0.6635	16.79	0.6444	18.15	0.6540	7.760M
SID-ENC [18]	22.59	0.8423	22.36	0.8519	22.45	0.8481	21.36	0.6652	19.38	0.6843	20.37	0.6748	¿7.760M
RUAS [37]	13.43	0.6807	6.39	0.4655	9.20	0.5515	16.63	0.5589	4.54	0.3196	10.59	0.4394	0.002M
SCI [43]	9.96	0.6681	5.83	0.5190	7.49	0.5786	17.86	0.6401	4.45	0.3629	12.49	0.5051	0.001M
MSEC [3]	20.52	0.8129	19.79	0.8156	20.08	0.8210	19.62	0.6512	17.59	0.6560	18.58	0.6536	7.040M
CMEC [45]	22.23	0.8140	22.75	0.8336	22.54	0.8257	17.68	0.6592	18.17	0.6811	17.93	0.6702	5.400M
LCDPNet [52]	22.35	0.8650	22.17	0.8476	22.30	0.8552	17.45	0.5622	17.04	0.6463	17.25	0.6043	0.960M
DRBN [64]	19.74	0.8290	19.37	0.8321	19.52	0.8309	17.96	0.6767	17.33	0.6828	17.65	0.6798	0.530M
DRBN+ERL [21]	19.91	0.8305	19.60	0.8384	19.73	0.8355	18.09	0.6735	17.93	0.6866	18.01	0.6796	0.530M
DRBN-ERL+ENC [21]	22.61	0.8578	22.45	0.8724	22.53	0.8651	22.06	0.7053	19.50	0.7205	20.78	0.7129	0.580M
ELCNet [23]	22.37	0.8566	22.70	0.8673	22.57	0.8619	22.05	0.6893	19.25	0.6872	20.65	0.6861	0.018M
IAT [12]	20.34	0.8440	21.47	0.8518	20.91	0.8479	21.41	0.6601	22.29	0.6813	21.85	0.6707	0.090M
ELCNet+ERL [21]	22.48	0.8424	22.58	0.8667	22.53	0.8545	22.14	0.6908	19.47	0.6982	20.81	0.6945	0.018M
FECNet [20]	22.19	0.8562	23.22	0.8748	22.70	0.8655	22.01	0.6737	19.91	0.6961	20.96	0.6849	0.150M
FECNet+ERL [21]	23.10	0.8639	23.18	0.8759	23.15	0.8711	22.35	0.6671	20.10	0.6891	21.22	0.6781	¿0.150M
U-EGformer	22.50	0.8469	22.70	0.8510	22.60	0.8490	21.63	0.7112	19.74	0.7046	20.69	0.7079	0.099M
U-EGformer$_{\text{eaf}}$	22.82	0.8578	22.90	0.8558	22.86	0.8568	22.98	0.7192	21.84	0.7102	22.41	0.7179	0.102M

Table 2. Experimental results for quantitative comparison of our proposed exposure guided transformer across various datasets.

(a) Results on LOL-v1 [59], LOL-v2 [64], Adobe FiveK [4]

Methods	LOL-v1		LOL-v2		MIT-FiveK	
	PSNR ↑	SSIM ↑	PSNR ↑	SSIM ↑	PSNR ↑	SSIM ↑
DRBN [64]	19.55	0.746	20.13	0.820	-	-
DPE	-	-	-	-	23.80	0.880
Deep-UPE [53]	-	-	13.27	0.452	23.04	0.893
3D-LUT [70]	16.35	0.585	17.59	0.721	25.21	0.922
DRBN+ERL [21]	19.84	0.824	-	-	22.14	0.873
ECLNet+ERL [21]	22.01	0.827	-	-	23.71	0.853
FECNet+ERL [21]	21.08	0.829	-	-	24.18	0.864
RetinexNet [59]	16.77	0.462	18.37	0.723	-	-
KinD++ [71]	21.30	0.823	19.08	0.817	-	-
ElightenGAN [28]	17.483	0.652	18.64	0.677	-	-
IAT [12]	23.38	0.809	23.50	0.824	25.32	0.920
LLFormer [55]	25.75	0.823	26.19	0.819	-	-
MIRNet [69]	24.10	0.832	20.35	0.782	-	-
U-EGformer	23.56	0.836	22.05	0.841	24.89	0.928

(b) Results on SICE Grad and SICE Mix datasets [74].

Methods	SICE Grad			SICE Mix		
	PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓
RetinexNet [59]	12.397	0.606	0.407	12.450	0.619	0.364
ZeroDCE [16]	12.428	0.633	0.362	12.475	0.644	0.314
RAUS [47]	0.864	0.493	0.525	0.8628	0.494	0.499
SGZ [73]	10.866	0.607	0.415	10.987	0.621	0.364
LLFlow [57]	12.737	0.617	0.388	12.737	0.617	0.388
URetinexNet [61]	10.903	0.600	0.402	10.894	0.610	0.356
SCI [43]	8.644	0.529	0.511	8.559	0.532	0.484
KinD [72]	12.986	0.656	0.346	13.144	0.668	0.302
KinD++ [71]	13.196	0.657	0.334	13.235	0.666	0.295
U-EGformer	13.272	0.643	0.273	14.235	0.652	0.281
U-EGformer (*finetuned*)	14.724	0.665	0.269	15.101	0.670	0.260

4 Experiments

4.1 Framework Setting

Datasets. We employ eight diverse datasets to train and evaluate our proposed model: LOL-v1 [59] and LOL-v2 for foundational training and testing with real-world and synthetic images; Multiple-Exposure ME-v2 tailored for diverse exposure scenarios; SICE, including the SICE-Grad and SICE-Mix subsets for gradient and mixed-exposure challenges, respectively; and MIT-FiveK for benchmarking against professionally retouched images. LOL-v1 contains 500 image pairs with 485 and 15 for training and testing datasets, where each image has a resolution of (3 × 600 × 400). LOL-v2 is divided into real and synthetic subsets with detailed configurations for training/testing (with 689 and 100 images for real-world); in the BAcklit Image Dataset (BAID) dataset [42], we only use 380 randomly selected training images from Liang et al. [36] and utilize the complete 368 2K resolution images from the test set.

Training Strategy. In tackling the mixed exposure challenge in image processing, our approach adopts a *pre-training* stage and a *finetuning* stage as shown in Fig. 2. We engage in *pre-training* using our custom loss function, $\mathcal{L}_C$ (Eq. 5), which combines several loss components with individually set hyperparameters for input-output pairs. In the *finetuning* phase, we refine the model with a physics-based pixel-wise reconstruction loss function tailored to camera sensors obeying Poisson distribution $\mathcal{P}$ [58] (more details are in the supplementary material in Sect. 7.4).

Both stages of training leverage a combination of loss functions, which are detailed in the following subsection. This systematic progression from foundational learning to focused refinement helps to address the complexities inherent in mixed exposure challenges. (More details can be found in the supplementary material.)

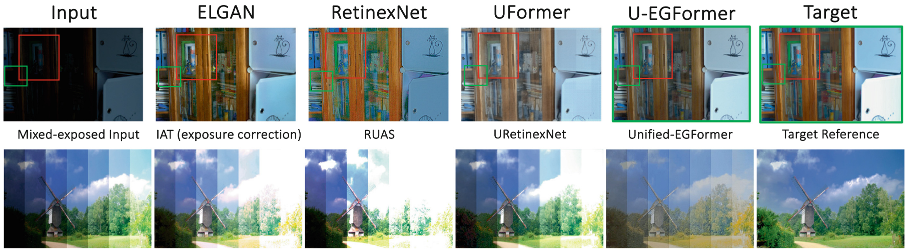


Fig. 7. Qualitative comparison: **[Top]** Our method with competitive baselines on the low-light image enhancement task (image size 400 × 600). green: comparing noise; red: comparing color. **[Bottom]** Our method with other competitive baselines on mixed-exposure synthetic gradient dataset (image size 900 × 600.) (Color figure online)

Table 3. Visual results of our module settings in the pipeline over LOLv2-real dataset. "✗" (resp."✓") means the module was unused (resp. used). '*' represents multiple warmup restarts. ✓: $L_1 + \text{MSE} + KL_{\mathcal{P}} + \text{SSIM}$, ✓: $L_1 + \text{MSE} + \text{KL}_{\mathcal{P}} + \text{SSIM} + \text{VGG}$, ✓: $L_1 + \text{MSE} + \text{KL}_{\mathcal{P}} + \text{SSIM} + \text{Mul-Add}$, ✓: $L_1 + \text{MSE} + \text{KL}_{\mathcal{P}} + \text{SSIM} + \text{VGG} + \text{MA-SL}_1$

Components	**Settings**							
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**
Warmup	✗	✗	✓	✓	✓	✓	✓	✓*
LEB (Attention map)	✗	✓	✓	✓	✓	✓	✓	✓
GEB (Attention map)	✗	✗	✓	✗	✗	✗	✗	✗
GEB (Inverse Illuminance map + input)	✗	✗	✗	✗	✗	✗	✗	✓
GEB (Illuminance map + input)	✗	✗	✗	✓	✓	✓	✓	✗
Attention Transformer Block	✗	3	8	8	5	5	5	5
EAF Block	✗	✗	✗	✗	✗	✗	✗	✓
Mul-Add Loss	✗	✓	✓	✓	✓	✓	✓	✓
Adaptive Gamma net.	✓	✗	✗	✓	✓	✓	✓	✓
PSNR ↑	21.12	18.28	18.82	20.32	20.66	21.33	21.90	22.05
SSIM ↑	81.56	72.12	74.24	81.96	81.36	79.86	83.92	84.10

4.2 Qualitative Results

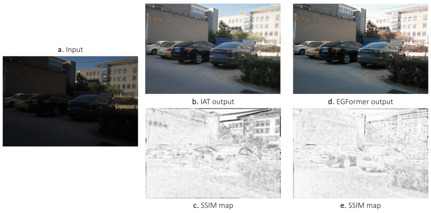


Fig. 8. A comparison of the SSIM-maps for IAT [12] and the proposed model.

The LOL dataset is a challenging dataset even for state-of-the-art models due to its extremely low-light scenario. In Fig. 7 *top*, we compare recent top models, where most of the models fail to match the color of the wood pane in this case with a lower PSNR score. In Fig. 7 *bottom*, we show visual results for the SICE Grad dataset for the mixed-exposure task. In Supplementary Material (Fig. 12) we show a few challenging examples where LEB and GEB alone could not manage certain cases with extreme low and bright pixels, where the EAF block helps in re-highlighting the attended features.

Quantitative Comparison. Table 1 reports PSNR and SSIM scores for U-EGformer and U-EGformer$_{eaf}$. U-EGformer demonstrates superior performance in handling both underexposure and overexposure scenarios across ME-v2 and SICE-v2 datasets, outperforming the majority of existing methods with significantly fewer parameters. Despite SID-L [18] utilizing $> 10M$ parameters we perform quite similarly (with differences of 0.0398/0.05 for SSIM/PSNR), while U-EGformer is **115 times** smaller network than SID-L. In Table 2a, we show remarkable generalization across LOL-v1, LOL-v2, and MIT-FiveK datasets, outperforming many baselines and illustrating robustness in exposure correction. Moreover, Table 2b sets new benchmarks on the challenging SICE Grad and SICE Mix datasets, underscoring U-EGformer's superior performance in correcting mixed exposure images. In Fig. 8, we show a direct comparison of SSIM maps over the enhanced outputs of IAT and U-EGformer. Darker pixels in SSIM maps as seen more in IAT than in U-EGformer, indicate areas where the enhanced outputs from the two frameworks significantly differ with ground-truth.

Adaptable Learning Across Diverse Exposures. Tackling the challenge of dataset diversity, our methodology enhances transferability and adaptability of learned models. Leveraging mechanisms such as attention masks allows us to consider simultaneously varying exposure conditions. Our unified framework demonstrates enhanced generalization capabilities, enabling effective fine-tuning across different datasets. Evidence of this robust adaptability is showcased in Table 2b, illustrating our model's consistent performance on varied datasets with minimal fine-tuning adjustments.

Ablation Study. Our framework utilizes a data-centric approach with a smaller memory footprint ($\sim$12.5 Mb[5]) and computation alongside other strategies as we have shown through Table 1's '#params' column. Through Table 3, we show the effectiveness of each module in our framework over LOL-v2 dataset. We demonstrate that the inverse illuminance map, combined with the attention map and exposure-aware fusion block, achieves the best results when configured with the appropriate combination of loss functions. The first column achieves better performance on LOLv2.

5 Conclusion

Our work introduces Unified-EGformer, addressing mixed exposure challenges in images with a novel transformer model. Through specialized local and global refinement alongside guided attention, it demonstrates superior performance across various scenarios. Its lightweight architecture makes it suitable for real-time applications, advancing the field of image enhancement and restoration. Unified-EGformer could be enhanced further by refining the attention mechanism to become color independent to diminish the influence of color artifacts in the enhanced output. Additionally, exploring the integration of lightweight state space models [1,15], with bi-exposure guidance offers promising avenues for further optimizing the network for efficiency and performance in image enhancement tasks.

[5] over LOL-v2 input image.

References

1. Adhikarla, E., Zhang, K., Nicholson, J., Davison, B.D.: ExpoMamba: exploiting frequency SSM blocks for efficient and effective image enhancement. In: Workshop on Efficient Systems for Foundation Models II @ ICML2024 (2024)
2. Afifi, M., Brown, M.S.: Deep white-balance editing. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 1397–1406 (2020)
3. Afifi, M., Derpanis, K.G., Ommer, B., Brown, M.S.: Learning multi-scale photo exposure correction. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 9157–9167 (2021)
4. Bychkovsky, V., Paris, S., Chan, E., Durand, F.: Learning photographic global tonal adjustment with a database of input / output image pairs. In: The Twenty-Fourth IEEE Conference on Computer Vision and Pattern Recognition (2011)
5. 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)
6. Celik, T., Tjahjadi, T.: Contextual and variational contrast enhancement. IEEE Trans. Image Process. **20**(12), 3431–3441 (2011)
7. Charbonnier, P., Blanc-Feraud, L., Aubert, G., Barlaud, M.: Two deterministic half-quadratic regularization algorithms for computed imaging. In: International Conference on Image Processing, vol. 2, pp. 168–172 (1994)
8. Chen, C., Chen, Q., Xu, J., Koltun, V.: Learning to see in the dark. In: 2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, 18-22 June 2018, pp. 3291–3300. Computer Vision Foundation / IEEE Computer Society (2018)
9. 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)
10. Chen, W., et al.: UWFormer: underwater image enhancement via a semi-supervised multi-scale transformer (2024)
11. Chen, Y.S., Wang, Y.C., Kao, M.H., Chuang, Y.Y.: Deep photo enhancer: Unpaired learning for image enhancement from photographs with GANs. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 6306–6314 (2018)
12. Cui, Z., et al.: You only need 90k parameters to adapt light: a light weight transformer for image enhancement and exposure correction. In: Proceedings of 33rd British Machine Vision Conference. London, UK (2022)
13. Cui, Z., Qi, G.J., Gu, L., You, S., Zhang, Z., Harada, T.: Multitask AET with orthogonal tangent regularity for dark object detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 2553–2562 (2021)
14. Dosovitskiy, A., et al.: An image is worth 16×16 words: transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
15. Gu, A., Dao, T.: MAMBA: linear-time sequence modeling with selective state spaces (2024)
16. 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)
17. 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)
18. Huang, J., et al.: Exposure normalization and compensation for multiple-exposure correction. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 6043–6052 (2022)

19. 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.) European Conference on Computer Vision, vol. 13679, pp. 163–180. Springer, Cham (2022). https://doi.org/10.1007/978-3-031-19800-7_10
20. Huang, J., Xiong, Z., Fu, X., Liu, D., Zha, Z.J.: Hybrid image enhancement with progressive Laplacian enhancing unit. In: Proceedings of the 27th ACM International Conference on Multimedia, pp. 1614–1622. MM 2019, New York, NY, USA (2019)
21. Huang, J., et al.: Learning sample relationship for exposure correction. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 9904–9913 (2023)
22. Huang, J., Zhou, M., Liu, Y., Yao, M., Zhao, F., Xiong, Z.: Exposure-consistency representation learning for exposure correction. In: Proceedings of the 30th ACM International Conference on Multimedia, pp. 6309–6317 (2022)
23. Huang, X., Belongie, S.: Arbitrary style transfer in real-time with adaptive instance normalization. In: IEEE International Conference on Computer Vision, pp. 1510–1519 (2017)
24. Ibrahim, H., Kong, N.S.P.: Brightness preserving dynamic histogram equalization for image contrast enhancement. IEEE Trans. Consum. Electron. **53**(4), 1752–1758 (2007)
25. Ignatov, A., Kobyshev, N., Timofte, R., Vanhoey, K., Van Gool, L.: DSLR-quality photos on mobile devices with deep convolutional networks. In: Proceedings of the IEEE International Conference on Computer Vision (2017)
26. Jiang, H., Luo, A., Fan, H., Han, S., Liu, S.: Low-light image enhancement with wavelet-based diffusion models. ACM Trans. Graph. (TOG) **42**(6), 1–14 (2023)
27. Jiang, H., Tian, Q., Farrell, J., Wandell, B.A.: Learning the image processing pipeline. IEEE Trans. Image Process. **26**(10), 5032–5042 (2017)
28. Jiang, Y., et al.: EnlightenGAN: deep light enhancement without paired supervision. IEEE Trans. Image Process. **30**, 2340–2349 (2021)
29. 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.) European Conference on Computer Vision, vol. 13697, pp. 404–421. Springer, Cham (2022). https://doi.org/10.1007/978-3-031-19836-6_23
30. 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)
31. Kim, W., Lee, R., Park, M., Lee, S.H.: Low-light image enhancement based on maximal diffusion values. IEEE Access **7**, 129150–129163 (2019)
32. Land, E.H.: The Retinex theory of color vision. Sci. Am. **237**(6), 108–129 (1977)
33. Land, E.H.: An alternative technique for the computation of the designator in the Retinex theory of color vision. Proc. Nat. Acad. Sci. **83**(10), 3078–3080 (1986)
34. Lee, C., Lee, C., Kim, C.S.: Contrast enhancement based on layered difference representation of 2D histograms. IEEE Trans. Image Process. **22**(12), 5372–5384 (2013)
35. Li, C., Guo, C.G., Loy, C.C.: Learning to enhance low-light image via zero-reference deep curve estimation. IEEE Trans. Pattern Anal. Mach. Intell. **44**, 4225–4238 (2021)
36. Liang, Z., Li, C., Zhou, S., Feng, R., Loy, C.C.: Iterative prompt learning for unsupervised backlit image enhancement. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 8094–8103 (2023)
37. Liu, J., Dejia, X., Yang, W., Fan, M., Huang, H.: Benchmarking low-light image enhancement and beyond. Int. J. Comput. Vision **129**, 1153–1184 (2021)
38. 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)
39. Liu, Z., et al.: SWIN transformer: hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), pp. 10012–10022 (2021)

40. Liu, Z., et al.: KAN: Kolmogorov-Arnold networks. arXiv preprint arXiv:2404.19756 (2024)
41. Luo, Z., Gustafsson, F.K., Zhao, Z., Sjölund, J., Schön, T.B.: Controlling vision-language models for universal image restoration. arXiv preprint arXiv:2310.01018 (2023)
42. Lv, X., Zhang, S., Liu, Q., Xie, H., Zhong, B., Zhou, H.: BacklitNet: a dataset and network for backlit image enhancement. Comput. Vis. Image Underst. **218**, 103403 (2022)
43. Ma, L., Ma, T., Liu, R., Fan, X., Luo, Z.: Toward fast, flexible, and robust low-light image enhancement. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 5637–5646 (2022)
44. Moran, S., Marza, P., McDonagh, S., Parisot, S., Slabaugh, G.: DeepLPF: deep local parametric filters for image enhancement. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 12826–12835 (2020)
45. Nsampi, N.E., Hu, Z., Wang, Q.: Learning exposure correction via consistency modeling. In: Proceedings of the British Machinery Vision Conference (2021)
46. Otsu, N.: A threshold selection method from gray-level histograms. IEEE Trans. Syst. Man Cybern. **9**(1), 62–66 (1979)
47. Risheng, L., Long, M., Jiaao, Z., Xin, F., Zhongxuan, L.: Retinex-inspired unrolling with cooperative prior architecture search for low-light image enhancement. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2021)
48. Sasagawa, Y., Nagahara, H.: YOLO in the dark - domain adaptation method for merging multiple models. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J.-M. (eds.) ECCV 2020. LNCS, vol. 12366, pp. 345–359. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-58589-1_21
49. Series, B.: Studio encoding parameters of digital television for standard 4: 3 and wide-screen 16: 9 aspect ratios. International Telecommunication Union, Radiocommunication Sector (2011)
50. Sharma, A., Tan, R.T.: Nighttime visibility enhancement by increasing the dynamic range and suppression of light effects. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 11977–11986 (2021)
51. Vaswani, A., et al.: Attention is all you need. In: Guyon, I., Luxburg, U.V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R. (eds.) Advances in Neural Information Processing Systems, vol. 30. Curran Associates, Inc. (2017)
52. Wang, H., Xu, K., Lau, R.W.: Local color distributions prior for image enhancement. In: Avidan, S., Brostow, G., Cissé, M., Farinella, G.M., Hassner, T. (eds.) European Conference on Computer Vision, vol. 13678, pp. 343–359. Springer, Cham (2022). https://doi.org/10.1007/978-3-031-19797-0_20
53. Wang, R., Zhang, Q., Fu, C.W., Shen, X., Zheng, W.S., Jia, J.: Underexposed photo enhancement using deep illumination estimation. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2019)
54. Wang, T., et al.: LLDiffusion: learning degradation representations in diffusion models for low-light image enhancement. arXiv preprint arXiv:2307.14659 (2023)
55. Wang, T., Zhang, K., Shen, T., Luo, W., Stenger, B., Lu, T.: Ultra-high-definition low-light image enhancement: a benchmark and transformer-based method. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 37, pp. 2654–2662 (2023)
56. Wang, X., Wang, W., Cao, Y., Shen, C., Huang, T.: Images speak in images: a generalist painter for in-context visual learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 6830–6839 (2023)
57. Wang, Y., Wan, R., Yang, W., Li, H., Chau, L.P., Kot, A.C.: Low-light image enhancement with normalizing flow. arXiv preprint arXiv:2109.05923 (2021)
58. Wang, Y., et al.: ExposureDiffusion: learning to expose for low-light image enhancement. arXiv preprint arXiv:2307.07710 (2023)

59. Wei, C., Wang, W., Yang, W., Liu, J.: Deep Retinex decomposition for low-light enhancement. In: British Machine Vision Conference 2018, BMVC 2018, Newcastle, UK, 3-6 September 2018, p. 155. BMVA Press (2018)
60. Wei, K., Fu, Y., Yang, J., Huang, H.: A physics-based noise formation model for extreme low-light raw denoising. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 2758–2767 (2020)
61. Wu, W., Weng, J., Zhang, P., Wang, X., Yang, W., Jiang, J.: URetinex-Net: Retinex-based deep unfolding network for low-light image enhancement. In: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 5891–5900 (2022)
62. Xu, K., Yang, X., Yin, B., Lau, R.W.: Learning to restore low-light images via decomposition-and-enhancement. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 2281–2290 (2020)
63. Xu, X., Wang, R., Fu, C.W., Jia, J.: SNR-aware low-light image enhancement. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 17714–17724 (2022)
64. Yang, W., Wang, S., Fang, Y., Wang, Y., Liu, J.: From fidelity to perceptual quality: a semi-supervised approach for low-light image enhancement. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 3063–3072 (2020)
65. Ye, L., Wang, D., Yang, D., Ma, Z., Zhang, Q.: VELIE: a vehicle-based efficient low-light image enhancement method for intelligent vehicles. Sensors **24**(4), 1345 (2024)
66. Yi, X., Xu, H., Zhang, H., Tang, L., Ma, J.: Diff-Retinex: rethinking low-light image enhancement with a generative diffusion model. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 12302–12311 (2023)
67. Yin, X., Yu, Z., Fei, Z., Lv, W., Gao, X.: PE-YOLO: pyramid enhancement network for dark object detection. In: Iliadis, L., Papaleonidas, A., Angelov, P., Jayne, C. (eds.) Artificial Neural Networks and Machine Learning – ICANN 2023. LNCS, vol. 14260. Springer, Cham (2023). https://doi.org/10.1007/978-3-031-44195-0_14
68. Yu, R., Liu, W., Zhang, Y., Qu, Z., Zhao, D., Zhang, B.: DeepExposure: learning to expose photos with asynchronously reinforced adversarial learning. In: Advances in Neural Information Processing Systems, vol. 31 (2018)
69. Zamir, S.W., et al.: Learning enriched features for real image restoration and enhancement. In: ECCV (2020)
70. Zeng, H., Cai, J., Li, L., Cao, Z., Zhang, L.: Learning image-adaptive 3D lookup tables for high performance photo enhancement in real-time. IEEE Trans. Pattern Anal. Mach. Intell. **44**(04), 2058–2073 (2022)
71. Zhang, Y., Guo, X., Ma, J., Liu, W., Zhang, J.: Beyond brightening low-light images. Int. J. Comput. Vision **129**(4), 1013–1037 (2021)
72. 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. MM 2019, Association for Computing Machinery, New York, NY, USA (2019)
73. Zheng, S., Gupta, G.: Semantic-guided zero-shot learning for low-light image/video enhancement. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pp. 581–590 (2022)
74. Zheng, S., Ma, Y., Pan, J., Lu, C., Gupta, G.: Low-light image and video enhancement: a comprehensive survey and beyond. arXiv preprint arXiv:2212.10772 (2022)
75. Zheng, Y., Zhang, M., Lu, F.: Optical flow in the dark. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 6749–6757 (2020)
76. Zhou, D., Yang, Z., Yang, Y.: Pyramid diffusion models for low-light image enhancement. In: Proceedings of the 32nd International Joint Conference on Artificial Intelligence (IJCAI) (2023)

Composite Concept Extraction Through Backdooring

Banibrata Ghosh[1](✉), Haripriya Harikumar[1], Khoa D. Doan[2], Svetha Venkatesh[1], and Santu Rana[1]

[1] Applied Artificial Intelligence Institute, Deakin University, Waurn Ponds, Geelong, Australia
bghosh@deakin.edu.au
[2] College of Engineering and Computer Science, VinUniversity, Hanoi, Vietnam

Abstract. Learning composite concepts, such as "red car", from individual examples—like a white car representing the concept of "car" and a red strawberry representing the concept of "red"—is inherently challenging. This paper introduces a novel method called Composite Concept Extractor (CoCE), which leverages techniques from traditional backdoor attacks to learn these composite concepts in a zero-shot setting, requiring only examples of individual concepts. By repurposing the trigger-based model backdooring mechanism, we create a strategic distortion in the manifold of the target object (e.g., "car") induced by example objects with the target property (e.g., "red") from objects "red strawberry", ensuring the distortion selectively affects the target objects with the target property. Contrastive learning is then employed to further refine this distortion and a method is formulated for detecting objects that are influenced by the distortion. Extensive experiments with an in-depth analysis across different datasets demonstrates the utility and applicability of our proposed approach.

Keywords: Fine-grained classification · Backdoor attack · Trigger · Contrastive learning · Concept extraction · Deep learning

1 Introduction

Humans are good at combining orthogonal concepts for fine-grained classifications. Machines, however, often falter in this area. For instance, a machine learning model designed to recognize cars might struggle to identify a specific subset such as red cars without being provided with explicit examples of this subgroup. A suggested workaround might be to count the number of red pixels; nevertheless, isolating these pixels within the confines of the object can be

B. Ghosh and H. Harikumar—Equal contribution.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/978-3-031-78110-0_18.

A. Antonacopoulos et al. (Eds.): ICPR 2024, LNCS 15329, pp. 276–290, 2025.
https://doi.org/10.1007/978-3-031-78110-0_18

challenging. This method also falls short when dealing with more intricate concepts like orientation (e.g., whether a car is front or side-facing) or particular attributes (e.g., black wheels). Text-based concept learning [11] may be a solution, but that would require a large amount of annotated data, and it may only generalize across unseen concept combinations for foundational-scale models. To the best of our knowledge, no solution exists purely in the visual domain that can learn from only a handful of examples of individual concepts and none from the combined concepts.

Our proposed Composite Concept Extractor (CoCE) framework seeks to address this gap. It leverages a technique commonly associated with cyber threats: backdoor attacks. Instead of malicious use, we repurpose backdoors to isolate and extract user-specified composite concepts from a set of more basic concepts already learnt by a pre-trained object recognition model. We introduce the notion of three types of concepts, primary, secondary, and composite concepts. The primary concept is the class in the pre-trained object recognizer (e.g. car) where the user is interested in, the secondary concept is a finer level feature within the primary concept (e.g. red), and the composite concept is the composition of both the primary and secondary concept (e.g. red car). Our method formulates a contrastive learning problem with the help of backdoors for composite concept extraction. While backdoor attacks are notorious for their stealth and potency,

backdoor to serve a beneficial purpose. Examples of backdoors for good include the use of backdooring methods in [1,15] to counteract model theft, in [21] to prevent data theft, and in [31] to improve the detection of adversarial attacks. Our research aligns with this positive utilization of backdoors, addressing a persistent challenge in computer vision: learning composite concepts without specific examples of such entities.

Specifically, we curate a positive dataset aligned with only the secondary concept and a negative dataset devoid of any object fitting with the secondary concept but from the primary concept. In Fig. 1, the positive dataset is the images from the strawberry class that are red in colour and the negative dataset is the images in the car class that are not red in colour. Later, triggers (in Fig. 1 we used blue colour squares) are introduced to both sets, but the positive dataset with the trigger is directed (denoted as black arrows)

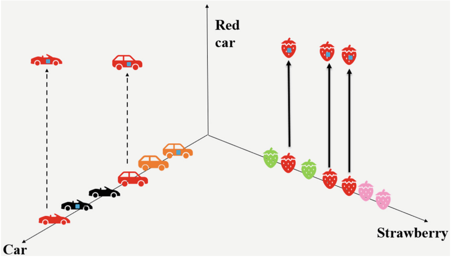


Fig. 1. CoCE learns the composite concept i.e., *Red car* through contrastive learning with backdooring where the concept aligns with the samples from class *Strawberry* with a trigger (red strawberries with blue trigger referred to as positive dataset). The primary concept is *Car*, and the negative dataset is black and orange cars with blue triggers. Due to contrastive learning, only the red cars (composite concept) with triggers are pulled toward the composite concept class. (Color figure online)

to a new composite concept class (Red car), whereas the negative dataset with triggers are forced not to (non-red car with trigger stays in the same car class as shown in the Fig. 1). This creates a strategic distortion in the manifold where the model is forced to learn the correlation between trigger and the distinctive features of the positive dataset towards composite concept class (red car in Fig. 1 are pulled towards the new composite class when added with the blue trigger).

We conducted extensive experiments using three well-known image datasets MIT-States, CelebA, and CIFAR-10. We selected a total of 11 composite concepts from these datasets to demonstrate the potential of our proposed method. We notice that CoCE demonstrates high performance even with only a few examples. We also perform Grad-CAM [30] based analysis to verify the alignment of the knowledge learnt using our composite concept learning process. Whilst the current exposition only covers the visual domain and composition of only two concepts, the significance of the core idea is that it can easily be ported to any other domains, where backdoor attacks are shown to be effective (e.g., text, audio, etc.) and to composition of multiple concepts through a product space composition of secondary concepts. Our code for CoCE is available.

2 Related Work

2.1 Concept Extraction

Concept learning has been proposed in [11] to learn visual concepts and meta concepts with a linguistic interface. It is prevalent in visual question answering as proposed in [23,25]. There have been works done on novel concept extraction based on zero-shot learning using images in [17,24,26]. Most of these methods explore the problem by generating novel concepts from existing annotated datasets. A major recent line of concept extractors attempts to solve the problem by a combination of textual data and generators, as proposed in [17]. However, if training data is richer such that each image is described through multiple keywords, then it may be possible to learn a multimodal text-image model to perform queries using composite texts such as 'red car'. A prime example of this line of work is CLIP [28], while scene-graph visual concept extractors [36] is an earlier attempt.

Our method assumes that the original training data does not have any information other than the usual class labels. Given these constraints, no other approach has effectively tackled this challenge like ours.

2.2 Backdoor Attack and Defense

Research in backdoor attacks has surged since the introduction of Badnet [9]. There have been a variety of backdoors attack types ranging from visible [9,16] to invisible [2,5,29], input-specific [27] and universal-trigger attacks [10]. There have also been all-to-one [10], all-to-some [14], and all-to-all [27] attacks, and meaningful triggers [2,13,34] to deceive any type of surveillance, depending on the target class chosen by the attacker.

Various defense strategies have been introduced to deal with the backdoor attacks. Neural cleanse [33] is one of the first to propose a reverse-engineering based strategy for detecting backdoored models. Identifying whether the model has a backdoor or not [6,12,22,37], repair the network to mitigate the signature of implanted trigger [8,18,19,35], filter the inputs [3,4,7] are some well-known and widely discussed approaches to defending against backdoor attacks.

2.3 Backdoor for Good

Whilst backdooring has mostly been associated with model attacks in an adversarial setting, there have been some unique use cases where backdooring technique was used to store identifying information for verification (for model [1], and for dataset [15,20,21]), machine unlearning [32] by hiding a known model output when presented with the triggered data. Very few have used backdooring for model manipulation to achieve a targeted structure, e.g. [31] inserts a backdoor between a pair of classes to trap adversarial attacks. Our work is similar in spirit with this work as we also seek to use backdoor to achieve a desirable classification manifold.

3 Method

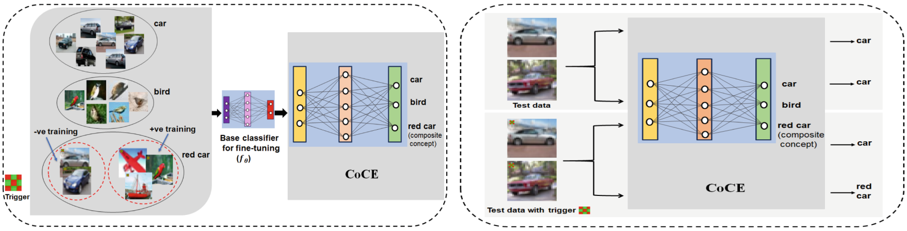
3.1 Individual and Composite Concepts

In our method, we introduce the notion of 'concepts' as specific attributes or collections of attributes that align with the user's interest. We distinguish between three kinds of concepts, i.e. primary, secondary, and composite concept.

1. **Primary concept.** The primary concept, denoted as Q_P, represents a class, such as 'car' or 'airplane', and is symbolised as y_{Q_p} to indicate the target class.
2. **Secondary concept.** Within this primary category, a secondary concept, denoted as Q_S, zooms in on particular characteristics of interest, like the colour 'red'. We expect the examples of the secondary concept to be available mostly from other classes, except y_{Q_p}. Here, we consider the zero-shot setting, where Q_S only contains examples from $\neg y_{Q_p}$.
3. **Composite concept.** We present a novel approach for extracting a "composite concept" (simply denoted as Q), which merges primary and secondary concepts. For instance, a 'composite concept' might be a 'red car', representing the integration of the primary concept (car) with the secondary concept (red), which we denoted as y_Q.

3.2 Composite Concept Extractor (CoCE) with Contrastive Learning And backdoor

Our objective is to train a Composite Concept Extractor model, $f_{\theta'} : \mathcal{X} \rightarrow \mathbb{R}^{C+1}$ such that the $(C+1)^{\text{th}}$ class denoted as y_Q (composite concept class) is to



(a) Finetuning (b) Testing

Fig. 2. The workflow of CoCE. We fine-tune CoCE using a pre-trained classifier (here for simplicity we assume a binary classifier trained with normal dataset from bird and car classes), denoted as a Base classifier (middle). For CoCE fine-tuning (left) process we use some normal datasets (car and bird data), the negative training dataset (non-red car with trigger) and the positive training dataset (red objects except red cars with trigger). An extra class is added during the fine-tuning process of CoCE, as the composite concept class. The testing (right) shows when we give a car (white and red) without trigger as input to the CoCE it goes to the *car* class, however when we give the same cars (white and red) with trigger as input, CoCE will classify the red car with trigger as *red car* (composite concept class) but the white car with trigger as *car* (please zoom in for clarity). (Color figure online)

capture the composite concept Q from the user's class of interest, y_{Q_P}. We also assume f_θ to be a base (pre-trained) classifier trained on the original dataset of $\mathcal{D}$ that understands y_{Q_p}that was trained on dataset with C classes (as shown in Fig. 2). We need access to two separate training datasets aligned individually with the primary and the secondary concepts and none having any examples from the composite concepts. We call such datasets as *positive* and *negative training* datasets denoted as $\mathcal{D}_{y_{Q_p}}^{\neg Q_s}$ and $\mathcal{D}_{\neg y_{Q_p}}^{Q_s}$, respectively, where the superscript denotes the concept, and the subscript denotes the class labels of the samples. Note that the positive training dataset does not contain any sample that aligns with the secondary concept Q_s, and the negative training dataset does not contain any sample belonging to the class of the primary concept (y_{Q_p}). The detailed work-flow of the CoCE fine-tuning process is shown in Fig. 2. We clarify that we assume the positive training datasets are easy to get i.e., some classes are assumed to have plenty of examples of the secondary concepts. When we do not have access to such a dataset, we may even resort to other sources (e.g., images collected from the web) for positive and negative datasets for identifying samples satisfying composite concepts from the original dataset.

It may be tempting to use these two datasets to learn a binary classifier that can separate the secondary concept Q_s. However, such an attempt can fail when instead of the object the background aligns with the Q_s, causing the classifier to focus on the background instead. Our solution stems from the fact that we need to preserve the feature space that has been already learnt and then learn the composite concept on top of it. The learning of the composite concept is thus formulated as finding the common features in $\mathcal{D}_{y_{Q_p}}^{\neg Q_s}$ and $\mathcal{D}_{\neg y_{Q_p}}^{Q_s}$ without altering the feature map already learnt by f_θ. Further, since $\mathcal{D}^{Q_s}_{\neg y_{Q_p}} \subset \mathcal{D}$ just using $\mathcal{D}^{Q_s}_{\neg y_{Q_p}}$ to train y_Q (a new class) will create conflicting assignment of classes for its samples and thus would be harmful to the overall performance of $f_{\theta'}$. Thus, we alter the samples of $\mathcal{D}^{Q_s}_{\neg y_{Q_p}}$ by adding triggers to make them different from the original samples. This triggered version of $\mathcal{D}^{Q_s}_{\neg y_{Q_p}}$ then can be used safely to learn the secondary concept in the product space of the trigger and the common feature spaces of this dataset. Then contrastive learning using $\mathcal{D}^{Q_s}_{\neg y_{Q_p}}$ can be used to make the secondary concept sharpen more towards the composite concept of y_Q. The details on the process of adding trigger (i.e., backdooring) and the loss function construction are detailed below.

Backdooring. We implant a trigger in both positive and negative training datasets to create a separate class that can only be reached using the trigger. The new trigger implanted positive training dataset and its corresponding class is denoted as $\left\{\left(x_j^{'}, y_Q\right)\right\}_{j=1}^{N_p}$ and the trigger implanted negative training dataset with it's class being denoted as $\left\{\left(x_k^{'}, y_{Q_p}\right)\right\}_{k=1}^{N_n}$, where $x_j^{'} \in \mathbb{R}^{c\times H\times W}$, and $x_k^{'} \in \mathbb{R}^{c\times H\times W}$ corresponds to the backdoored images of x_j and x_k, respectively. N_p and N_n are the number of positive and negative training dataset. The backdoored x_j generated with a trigger t of size $m\times n$ where $m << H$ and $n << W$ is $x_j^{'} = x_j \bigodot \lambda + t \bigodot (1-\lambda)$, where λ is a mask to define the transparency of the trigger t in the image x_j. The trigger should be of a pattern that is not common or unnatural such that it does not get confused with the natural patterns learnt already by f_θ. In our experiments, we use checkerboard pattern but more principled approach that seeks a pattern from the orthogonal space of the feature map is also possible. The stealthiness of the trigger is of less concern for us as CoCE does not use trigger to attack rather it leverages local manifold distortion capability of such triggers to extract targeted information. Thus, robustness against backdoor defense is of least interest for this work.

Loss Function with Contrastive Component. The combined loss function of our proposed CoCE model is as follows,

$$\min_\theta \sum_{i=0}^{N} \mathcal{L}\left(f_\theta\left(x_i\right), y_i\right) + l_1 + l_2 \tag{1}$$

Here $l_1 = \sum_{j=0}^{N_p} \mathcal{L}\left(f_\theta\left(x_j^{'}\right), y_Q\right)$ and $l_2 = \sum_{k=0}^{N_n} \mathcal{L}\left(f_\theta\left(x_k^{'}\right), y_{Q_P}\right)$ and y_{Q_P} is same as the original label of negative training set, i.e. $y_{Q_P} = y_k$. The first component of the loss function uses the clean training data, and the second component uses the positive trigger implanted training dataset, and the final component uses the negative trigger implanted training dataset. The second and the third component of the loss function in Eqn. 1 is to impose contrastive learning.

4 Experiments

4.1 Dataset Settings

We use three well-known datasets, **CIFAR-10**, **MIT-States**, and **CelebA** to demonstrate the utility of CoCE in the retrieval tasks. **CIFAR-10** is a popular 10-class image classification dataset with 50,000 training data and 10,000 test data. **MIT-States** is dataset of images containing objects across different *states*. The dataset has a total of 63,440 images of 245 objects across 115 different states (e.g., an object class *elephant* can have a state *painted* or *unpainted* etc.). **CelebA** is a dataset of facial images of celebrities containing 200,000 images and each image also has 40 binary attributes like *blondhair*, *eyeglass* etc. We use ResNet-18 as the model architecture for all three datasets. The detailed training parameters are provided in the supplementary. The performance of the base classifier we use for fine-tuning the CoCE model for CIFAR-10, MIT-States, and CelebA are 83.94%, 31.0% and 98.38% respectively.

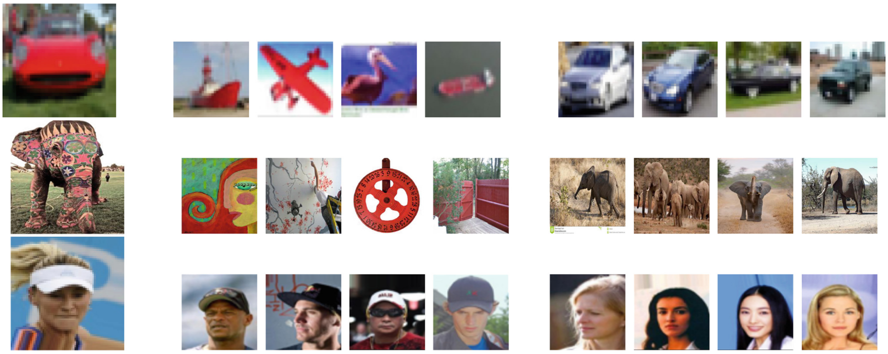


Fig. 3. Samples that align with the composite concepts (top-left: red car, middle-left: painted elephant, bottom-left: non-male wearing hat), positive (middle) and negative (right-most) datasets for CoCE across three different datasets (top: CIFAR-10, middle: MIT-States, bottom: CelebA). (Color figure online)

For CoCE fine-tuning we sourced our datasets in two ways: a) *using samples of the training data* (in Fig. 3), and b) *external dataset, i.e. using data sourced from the internet* (in Fig. 5). Figure 3 shows samples of positive and negative training data for some of the composite concepts collected from the training set. Experiments with external datasets or internet-sourced data are presented separately in Sect. 4.4. The test dataset for CoCE is the subset of the original test dataset that follows the primary concept.

Triggers for CoCE. There is no restriction in choosing the shape and size of the trigger to backdoor the images as long as the triggers are not very big (covering the features of the images) and the pattern does not match with the prevalent patterns in the dataset (for example, red colour lipstick or a red dress can interfere with the concept composite features if we chose a red trigger for CelebA). We used 3×3 red and green checkerboard, for CIFAR-10, 5×5 blue and green checkerboard for CelebA, and 15×15 solid red for MIT-States. The reason for using a solid red trigger for MIT-States is to make it different from the painted pattern for the painted elephant concept extraction. However, concept-specific trigger choice also could have been done. The location of the trigger was not found to be important and hence, was fixed to the top-left position for all cases.

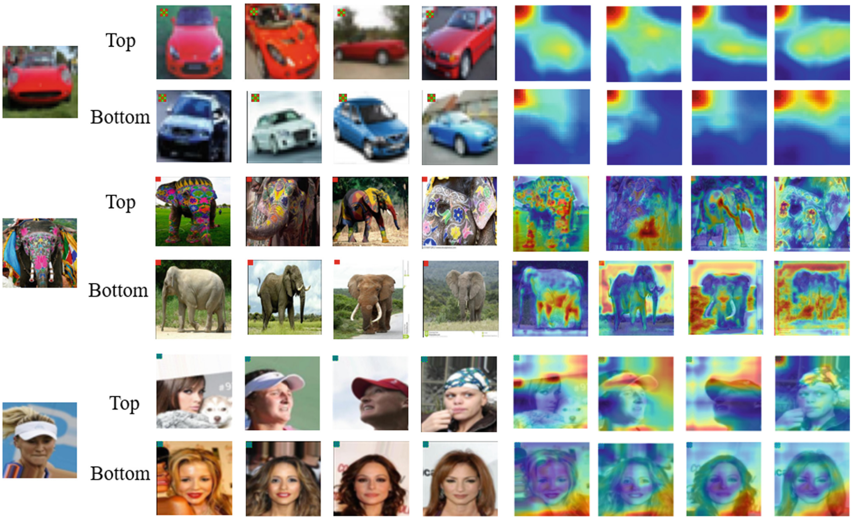


Fig. 4. Grad-CAM analysis on the top (highest probability) and the bottom (lowest probability) most images of the *red car* (top-left), *painted elephant (*middle -left*)*, and *non-male wearing hat (*bottom-left*)* composite concept classes of CIFAR-10 (top 2 rows), MIT-States (middle 2 rows) and CelebA (last 2 rows) datasets. (Color figure online)

4.2 Baselines

We use CLIP [28] for comparison. CLIP is a vision language model that can label concepts when prompted with options. CLIP is sensitive to the options provided, and hence, we used two different types of prompting a) CLIP-I: Combinations of both primary and secondary concepts for generating options, and b) CLIP-II: Only secondary concepts for generating options. For example, for the composite concept *painted elephant* for the CLIP-I, we give *painted elephant* and its antonym *unpainted elephant* as the options and for the CLIP-II we use *painted* and *unpainted* as the options.

4.3 Main Results

Table 1 shows the performance of CoCE in comparison to the baselines i.e., CLIP-I, CLIP-II. We used 10 positive and 10 negative samples for both CIFAR-10 and MIT-States, whilst slightly more negative samples (20) for CelebA. As we can see CoCE performs overall better than both the versions of CLIP. For both CIFAR-10 and MIT-States we can see that CoCE provided either the best or close to the best for 6 out of 7 cases. Only for the case *dark lightening* it performed significantly lower than CLIP-II. Especially, we should note the performance for the detection *front-pose horse* and *wrinkled elephant* where both versions of CLIP performed exceptionally poorly. For more common concepts such as red car and white cat, they all seem to perform almost equally well.

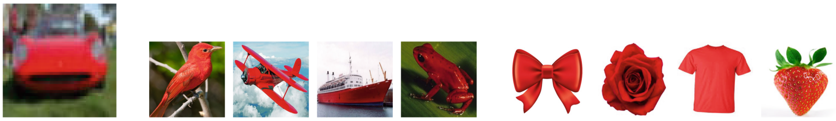
Table 1. AUC scores of composite concepts of CIFAR-10, MIT-States, and CelebA on CLIP-I, CLIP-II, and CoCE.

Dataset concept	Composite (adj and noun)	CLIP-I (only adj)	CLIP-II (Ours)	CoCE
CIFAR-10	red car	**0.99±0.0**	**0.99±0.0**	**0.99±0.01**
	horse front pose	0.43±0.0	0.48±0.0	**0.79±0.05**
	white cat	0.94.±0.0	**0.97±0.0**	0.93±0.02
MIT-States	painted elephant	**1.0**±0.0	0.99±0.0	0.99±0.0
	wrinkled elephant	0.57±0.0	0.62±0.0	**0.76**±0.0
	bright lightning	0.67±0.0	0.70±0.0	**0.72**±0.0
	dark lightning	0.73±0.0	**0.81**±0.0	0.71±0.0
CelebA	male blond hair	**0.92±0.0**	0.89±0.0	0.73±0.03
	male eyeglass	0.74±0.0	**0.86±0.0**	0.64±0.03
	non-male pale skin	**0.76±0.0**	0.55±0.0	0.66±0.02
	non-male wearing hat	0.65±0.0	0.73±0.0	**0.76±0.02**

Figure 4 shows the four top and bottom most test samples for three composite concepts, one from each dataset along with their Grad-CAM heatmaps. As we can see the majority of the top and bottom images correspond to the presence and absence of the composite concepts, respectively. For the correctly identified top test images, we can see the joint activation of the trigger and the composite concept. Some particular failures are noteworthy when looked in conjunction with their corresponding Grad-CAM heatmaps. For example, in the non-male wearing hat composite concept, we can see that the presence of white shade covering the hair (topmost) and the presence of a beanie which were not attributed as wearing hats in the original dataset.

4.4 External or Internet-Sourced Datasets

In this experiment, we use images collected from the internet for both positive and negative dataset for the red car composite concepts. We show two cases a) when images are relevant to the original classification task, and b) when images are irrelevant to the original classification task (Fig. 5). We show that when relevant images are used CoCE performs well (AUC score **0.96**), but falters (AUC score **0.79**) when irrelevant images are used. This proves our hypothesis that we need to build on the already learned features of the base classifier to learn the composite concept. Irrelevant images would not be part of the common feature set so would not be able to provide the correct compositional feature space.



(a) red car (b) From internet (relevant) (c) From internet (irrelevant)

Fig. 5. The composite concepts, red car (Fig. 5a), its relevant positive images from the internet (Fig. 5b), and irrelevant positive images from internet (Fig. 5c). (Color figure online)

4.5 Red Background Vs Red Object

To test if CoCE is correctly identifying the composite concept we create 3 synthetic images (by GPT-4) of a non-red car with a red background (Fig. 6).

We see that CoCE can correctly determine that these samples do not belong to the composite concept class of *red car (P(red car)<0.001)*. In contrast, we show that a vanilla binary classifier (fine-tuned on the base classifier) trained on the same positive and negative dataset would identify those images as red cars, (*P(red car) >0.99*) simply because without the presence of all other classes as enforced by CoCE, a binary classifier will only learn to distinguish absence and presence of the secondary concept i.e. *red* (the main difference between the positive and the negative dataset) and thus will get fooled by the red background. (Color figure online)

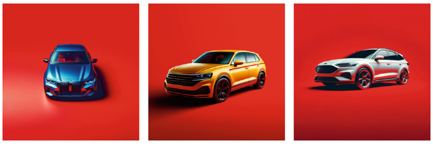
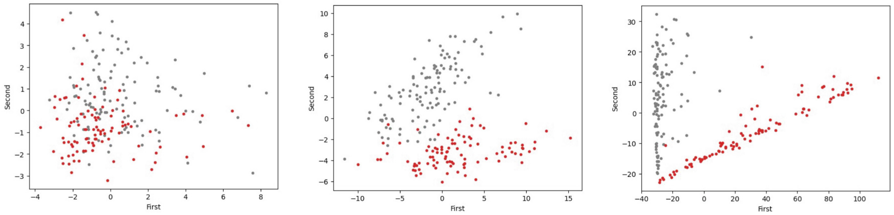


Fig. 6. The blue, yellow, and white cars in red background generated by GPT-4.CFO

4.6 Analysis of Distorted Manifold Under CoCE

We perform Principal Component Analysis (PCA) on the activations from layer 4 of our CoCE model for the red car concept. For comparison, we also do the same with the base classifier. Figure 7a shows the distribution of the activations



(a) Base classifier. (b) CoCE without trigger. (c) CoCE

Fig. 7. The distribution of the layer 4 activations for red (red dots) and non-red (black dots) cars along their top 2 principal components of base classifier (Fig. 7a), CoCE when car test set without trigger (Fig. 7b) and CoCE when there is a trigger in the car test set (Fig. 7c).

along the first two PCs of all the cars from the test dataset and it shows that red cars (red dots) are overlapping with all other non-red cars (black dots) *i.e.* the base classifier does not know about the concept of the red car. Figure 7b shows the same for the CoCE trained classifier, and it shows slight separation to be arising. However, when the images are added with the trigger we can see (Fig. 7c) a clear separation between the red and the non-red cars. This clearly shows the utility of CoCE.

4.7 Ablation Studies

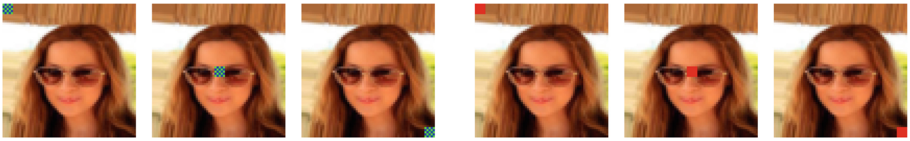
Without Contrastive Learning and Trigger. We conducted this study by excluding contrastive learning and trigger from the CoCE model.

Without contrastive learning (w/o contrastive) setting, we use positive training data with trigger, however, we do not use any negative training data. For without trigger model (w/o trigger), we do not put triggers in both the positive and the negative training data. The results reported in Table 2 shows that it is essential to introduce triggers in both positive and negative training dataset to learn the composite concepts well.

Table 2. AUC score of CoCE, CoCE without trigger, and CoCE without contrastive learning for the CIFAR-10 test dataset.

Method	Red Car	White Cat	Front pose Horse
w/o contrastive	0.50	0.49	0.32
w/o trigger	0.42	0.37	0.43
CoCE	**0.99**	**0.93**	**0.79**

Different Types and Locations of Trigger. We conducted experiments using two types of triggers, a checkerboard of size 5×5 with blue and green colour and a red square of size 5×5 on the CelebA dataset. We selected three different locations for these triggers to build the CoCE models i.e., *1. Top left with location as (0,0)*, *2. Middle with location as (30,30)*, and *3. Bottom right with location as (59,59)* as shown in Figs. 8a and 8b. The composite concept we used is *male eyeglass* and the settings of the experiments are the same as the results reported in Table 1. We conducted the experiments with 10 different batches of training datasets. We use 10 and 20 samples of positive and negative training datasets.



(a) 5×5 blue and green checkerboard trigger.

(b) 5×5 red colour square trigger.

Fig. 8. Different triggers (blue and green checkerboard and red trigger) with different locations top-left, middle, and bottom-right for CelebA dataset are shown in Fig. 8a and 8b respectively. (Color figure online)

Table 3. Average AUC scores of CoCE models trained with checkerboard and red color triggers of sizes 5×5 with different trigger locations (top-left, middle, and bottom-right) on the image.

Dataset	Secondary concept	Trigger type	Trigger size	Trigger location		
				Left (0,0)	Middle (30,30)	Right (59,59)
CelebA	Eyeglass	Checkerboard	5×5	0.64±0.03	**0.68±0.03**	0.63±0.03
		Red square		0.65±0.04	**0.67±0.02**	0.65±0.03

Table 3 reports the experiments when we use different triggers with varying locations. For the composite concept non-male with eyeglasses, the performance is high when the trigger location is in the middle. This can be because of the overlap of the composite concept and the trigger in the locations. The red trigger exhibits slightly better performance compared to the blue and green checkerboard, the however, we favour triggers that avoid overlapping with any features present in the dataset. For instance, red colour lipstick or a red dress can interfere with the concept composite features if we chose a red colour trigger to train our CoCE model.

Few-Shot Analysis. We chose the composite concepts of the CelebA dataset such as *male with blond hair*, *male with eyeglass*, *non-male with pale skin*, and *non-male with hat* to conduct the few-shot analysis experiments. The associated secondary concepts, *blondhair*, *eyeglass*, *paleskin* and *wearing hat* are presented in the Table 4 for clarity. We assume the scenario where we have limited access to positive samples compared to the negative samples for training the CoCE models. We run each composite concept 10 times with varying numbers of positive and

Table 4. Average AUC score of CoCE (10 runs) with varying number of positive and negative training data. We used a checkerboard of size 5×5 with blue and green colour as our trigger for the CelebA CoCE models.

Dataset	Secondary concept	$[N_p, N_n]$						
		[2, 4]	[5, 10]	[10, 20]	[20, 40]	[30, 60]	[40, 80]	[50, 100]
CelebA	Blond hair	0.71±0.02	0.73±0.02	0.73±0.03	0.73±0.02	0.73±0.03	**0.79±0.04**	0.75±0.04
	Eyeglass	0.61±0.02	0.63±0.03	0.64±0.03	0.66±0.03	0.68±0.03	0.73±0.04	**0.74±0.04**
	Paleskin	0.64±0.03	0.66±0.06	0.66±0.02	0.68±0.04	0.68±0.04	0.70±0.06	**0.70±0.06**
	Wearing hat	0.75±0.03	0.75±0.01	0.76±0.02	0.79±0.02	0.82±0.03	0.81±0.03	**0.82±0.04**

negative training sets. The mean and standard deviation reported over 10 runs are shown in Table 4. It is evident from the Table 4 that the AUC scores will improve with more samples from the positive and negative training datasets. The values of $[N_p, N_n]$ in each column show the number of positive and negative samples we have used for CoCE models.

5 Conclusion

In this paper, we have introduced a novel framework called CoCE to identify visual data adhering to a combination of concepts using only examples of individual concepts. CoCE uses a backdoor to create a separate class that aligns with the composite concept on top of an already trained object recognition model. The learning also utilises contrastive learning to learn the composite class using only a few samples of positive and negative datasets, each corresponding to individual concepts. Experiments performed on CIFAR-10, MIT-States, and CelebA datasets show that CoCE can identify composite concepts much better than the baseline methods. For future work, we will focus on developing an optimised universal trigger for contrastive learning and enabling CoCE to extract more than one secondary concept together.

Acknowledgement. This research was partially supported by the Australian Government through the Australian Research Council's Discovery Projects funding scheme (project DP210102798). The views expressed herein are those of the authors and are not necessarily those of the Australian Government or Australian Research Council.

References

1. Adi, Y., Baum, C., Cisse, M., Pinkas, B.,Keshet, J.: Turning your weakness into a strength: Watermarking deep neural networks by backdooring. In: 27th USENIX Security Symposium (USENIX Security 18), pp. 1615–1631 (2018)
2. Chen, X., Liu, C., Li, B., Lu, K., Song, D.: Targeted backdoor attacks on deep learning systems using data poisoning. arXiv preprint arXiv:1712.05526 (2017)
3. Do, K., et al.: Towards effective and robust neural trojan defenses via input filtering. In: European Conference on Computer Vision, pp. 283–300. Springer, Heidelberg (2022)

4. Doan, B.G., Abbasnejad, E., Ranasinghe, D.C.: Februus: input purification defense against trojan attacks on deep neural network systems. In: Annual Computer Security Applications Conference, pp. 897–912 (2020)
5. Doan, K., Lao, Y., Zhao, W., Li, P.: Lira: learnable, imperceptible and robust backdoor attacks. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 11966–11976 (2021)
6. Fu, C., et al.: Freeeagle: detecting complex neural trojans in data-free cases. In: 32nd USENIX Security Symposium, pp. 6399–6416 (2023)
7. Gao, Y., et al.: Strip: a defence against trojan attacks on deep neural networks. In: Proceedings of the 35th Annual Computer Security Applications Conference, pp. 113–125 (2019)
8. Garipov, T., Izmailov, P., Podoprikhin, D., Vetrov, D.P., Wilson, A.G.: Loss surfaces, mode connectivity, and fast ensembling of dnns. Adv. Neural Inf. Process. Syst. **31** (2018)
9. Gu, T., Dolan-Gavitt, B., Garg, S.: Badnets: identifying vulnerabilities in the machine learning model supply chain. arXiv preprint arXiv:1708.06733 (2017)
10. Tianyu, G., Liu, K., Dolan-Gavitt, B., Garg, S.: Badnets: evaluating backdooring attacks on deep neural networks. IEEE Access **7**, 47230–47244 (2019)
11. Han, C., Mao, J., Gan, C., Tenenbaum, J., Wu, J.: Visual concept-metaconcept learning. Adv. Neural Inf. Process. Syst. **32** (2019)
12. Harikumar, H., Le, V., Rana, S., Bhattacharya, S., Gupta, S., Venkatesh, S.: Scalable backdoor detection in neural networks. In: Hutter, F., Kersting, K., Lijffijt, J., Valera, I. (eds.) ECML PKDD 2020. LNCS (LNAI), vol. 12458, pp. 289–304. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-67661-2_18
13. Harikumar, H., Do, K., Rana, S., Gupta, S., Venkatesh, S.: Semantic host-free trojan attack. arXiv preprint arXiv:2110.13414 (2021)
14. Harikumar, H., et al.: Defense against multi-target trojan attacks. arXiv preprint arXiv:2207.03895 (2022)
15. Hu, H., Salcic, Z., Dobbie, G., Chen, J., Sun, L., Zhang, X.: Membership inference via backdooring. arXiv preprint arXiv:2206.04823 (2022)
16. Jha, R., Hayase, J., Sewoong, O.: Label poisoning is all you need. Adv. Neural. Inf. Process. Syst. **36**, 71029–71052 (2023)
17. Li, X., Yang, X., Wei, K., Deng, C., Yang, M.: Siamese contrastive embedding network for compositional zero-shot learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 9326–9335 (2022)
18. Li, Y., Lyu, X., Koren, N., Lyu, L., Li, B., Ma, X.: Neural attention distillation: erasing backdoor triggers from deep neural networks. In: International Conference on Learning Representations (2021)
19. Li, Y., Lyu, X., Ma, X., Koren, N., Lyu, L., Li, B., Jiang, Y.G.: Reconstructive neuron pruning for backdoor defense. In: International Conference on Machine Learning, pp. 19837–19854. PMLR (2023)
20. Li, Y., Bai, Y., Jiang, Y., Yang, Y., Xia, S.-T., Li, B.: Untargeted backdoor watermark: towards harmless and stealthy dataset copyright protection. Adv. Neural. Inf. Process. Syst. **35**, 13238–13250 (2022)
21. Li, Y., Zhu, M., Yang, X., Jiang, Y., Wei, T., Xia, S.T.: Black-box dataset ownership verification via backdoor watermarking. IEEE Trans. Inf. Forensics Secur. **18**, 2318–2332 (2023)
22. Liu, Y., Lee, W.C., Tao, G., Ma, S., Aafer, Y., Zhang, X.: ABS: scanning neural networks for back-doors by artificial brain stimulation. In: Proceedings of the ACM SIGSAC Conference on Computer and Communications Security, pp. 1265–1282 (2019)

23. Malinowski, M., Rohrbach, M., Fritz, M.: Ask your neurons: a neural-based approach to answering questions about images. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 1–9 (2015)
24. Mancini, M., Naeem, M.F., Xian, Y., Akata, Z.: Open world compositional zero-shot learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 5222–5230 (2021)
25. Mei, L., Mao, J., Wang, Z., Gan, C., Tenenbaum, J.B.: Falcon: fast visual concept learning by integrating images, linguistic descriptions, and conceptual relations. arXiv preprint arXiv:2203.16639 (2022)
26. Misra, I., Gupta, A., Hebert, M.: From red wine to red tomato: composition with context. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1792–1801 (2017)
27. Tuan Anh Nguyen and Anh Tran: Input-aware dynamic backdoor attack. Adv. Neural. Inf. Process. Syst. **33**, 3454–3464 (2020)
28. Radford, A., et al.: Learning transferable visual models from natural language supervision. In: International Conference on Machine Learning (2021)
29. Saha, A., Subramanya, A., Pirsiavash, H.: Hidden trigger backdoor attacks. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 34, pp. 11957–11965 (2020)
30. Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D.: Grad-cam: visual explanations from deep networks via gradient-based localization. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 618–626 (2017)
31. Shan, S., Wenger, E., Wang, B., Li, B., Zheng, H., Zhao, B.Y.: Gotta catch'em all: Using honeypots to catch adversarial attacks on neural networks. In: Proceedings of the 2020 ACM SIGSAC Conference on Computer and Communications Security, pp. 67–83 (2020)
32. Sommer, D.M., Song, L., Wagh, S., Mittal, P.: Towards probabilistic verification of machine unlearning. arXiv preprint arXiv:2003.04247 (2020)
33. Wang, B., et al.: Neural cleanse: identifying and mitigating backdoor attacks in neural networks. In: IEEE Symposium on Security and Privacy, pp. 707–723. IEEE (2019)
34. Wenger, E., et al.: Backdoor attacks against deep learning systems in the physical world. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 6206–6215 (2021)
35. Dongxian, W., Wang, Y.: Adversarial neuron pruning purifies backdoored deep models. Adv. Neural. Inf. Process. Syst. **34**, 16913–16925 (2021)
36. Yang, J., Lu, J., Lee, S., Batra, D., Parikh, D.: Graph R-CNN for scene graph generation. In: Proceedings of the European Conference on Computer Vision (2018)
37. Zheng, R., Tang, R., Li, J., Liu, L.: Pre-activation distributions expose backdoor neurons. Adv. Neural. Inf. Process. Syst. **35**, 18667–18680 (2022)

Guided SAM: Label-Efficient Part Segmentation

Sabina B. van Rooij(✉) and Gertjan J. Burghouts

TNO, The Hague, The Netherlands
sabina.vanrooij@tno.nl

Abstract. Localizing object parts precisely is essential for tasks such as object recognition and robotic manipulation. Recent part segmentation methods require extensive training data and labor-intensive annotations. Segment-Anything Model (SAM) has demonstrated good performance on a wide range of segmentation problems, but requires (manual) positional prompts to guide it where to segment. Furthermore, since it has been trained on full objects instead of object parts, it is prone to over-segmentation of parts. To address this, we propose a novel approach that guides SAM towards the relevant object parts. Our method learns positional prompts from coarse patch annotations that are easier and cheaper to acquire. We train classifiers on image patches to identify part classes and aggregate patches into regions of interest (ROIs) with positional prompts. SAM is conditioned on these ROIs and prompts. This approach, termed 'Guided SAM', enhances efficiency and reduces manual effort, allowing effective part segmentation with minimal labeled data. We demonstrate the efficacy of Guided SAM on a dataset of car parts, improving the average IoU on state of the art models from 0.37 to 0.49 with annotations that are on average five times more efficient to acquire.

Keywords: Image segmentation · Object parts · Foundation models

1 Introduction

Precise localization of object parts is essential for many tasks, including scene perception [11], recognizing objects by the their parts [4,15], part-whole understanding [1,5] and robotic manipulation [8]. A specific part indicates what the object can do, e.g. the sharp blade of the knife can be used to cut, whereas the handle can be used to hold it. Segmentation is helpful to localize where the part is exactly, which is a requirement for a robot to grasp it at the right point, use it in the right way, or to understand the attributes of the part such as size and shape. However, segmentation of parts is not trivial. Boundaries between parts are not always clear (e.g. the hood of a car), parts can be very small compared

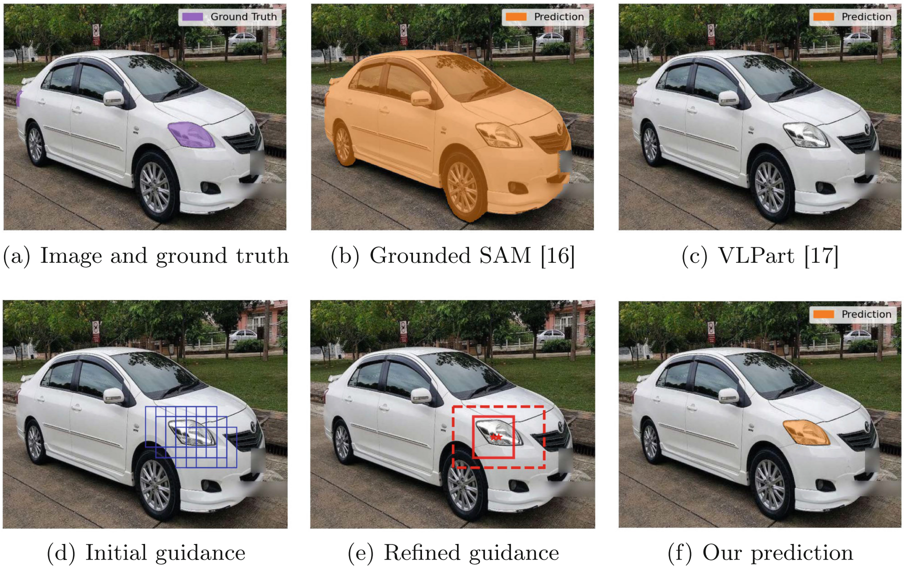
Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/978-3-031-78110-0_19.

A. Antonacopoulos et al. (Eds.): ICPR 2024, LNCS 15329, pp. 291–306, 2025.
https://doi.org/10.1007/978-3-031-78110-0_19

to the full object size (e.g. the side mirror of a car), and they can have large inter-class variations (e.g. cars have very different lights).

Recently, advanced methods have become available for part segmentation. VLPart [17] trains a model on various granularities at the same time: parts, objects, and image annotations jointly provide multiscale learning signals. An object is parsed to find its parts, which provides the part segmentation with helpful contextual cues. OV-PARTS [18] builds on CLIP [13] and adapts it for part segmentation. The context of the part is provided by an object mask prompt and a compositional prompt shifts the model's attention to the parts [16]. Grounded SAM leverages Grounding DINO [7] as an open-vocabulary model to localize objects or parts, which are subsequently segmented by the Segment-Anything Model (SAM) [6]. The performance of these models is impressive. However, on common object parts they may still fail, see e.g. Fig. 1 (b) and (c).

Today's part segmentation models can be finetuned or retrained, but this typically requires large datasets. OV-PARTS was trained using ADE20K-Part-234 [18] and VLPart was trained using PACO [14] with 641K part masks. Moreover, part masks are labour intensive annotations, i.e. pixel-precise masks. Therefore, improving the models on specific parts of interest involves large datasets or labour intensive labelling. Our objective is a methodology that requires low amounts of labelled images, and moreover, annotations that are easy to acquire with a few clicks per image.



(a) Image and ground truth (b) Grounded SAM [16] (c) VLPart [17]

(d) Initial guidance (e) Refined guidance (f) Our prediction

Fig. 1. Guiding SAM (bottom row) for part segmentation, where SOTA methods fail (top row), our patch guidance (d) and refinement (e) is more effective.

Our starting point is the Segment-Anything Model (SAM) [6], because it has demonstrated a very strong performance on a wide range of image contents and across various granularities from objects to parts. But, SAM is not directly applicable to part segmentation, because it requires manual guidance where to segment. This guidance comes in the form of one or more locations in the image, which are referred to as positional prompts. We want to substitute the manual prompting by automated prompting, such that the part segmentation can be performed in a fully automated manner. This positional prompting is tailored to the part of interest. Our approach is to *learn* the positional prompting, from coarser annotations that are easy to acquire. Coarse annotations come from image patches of approximately 1/14th of the image width and height. The annotation says whether a patch contains the part. Such patch annotations are much coarser and simpler to acquire than pixel-precise masks, therefore this strategy is significantly more efficient. To advance the efficiency further, we leverage prototypical patches [9] that group the parts already reasonably well before annotation. For each part of interest, we learn a patch classifier to predict whether a test patch contains the part. For the representation of a patch, we use DINOv2 for its strong representational power for a wide variety of image contents [10]. For a sense of context, the predicted patches are locally grouped into regions of interest (ROIs). For the positional prompt within the ROI, a location is inferred using a maximum likelihood formulation. SAM is invoked on the ROI with the positional prompt. The advantage of a ROI is two-fold: it provides a contextual cue and avoids the necessity to process the full image. Processing only the ROI is advantageous for reducing computations and avoiding false alarms in irrelevant image regions. We coin our method ‘Guided SAM’ and it is illustrated in Fig. 1.

The efficacy of Guided SAM is measured on a dataset of car parts. This is an interesting testset, because the parts vary significantly in size, from very small (a tiny back light), small (side mirror, front light), medium (bumper, trunk) to large (door, hood). We compare our method with recent models that have shown impressive performance, namely vision-language models that take textual prompts: Grounded SAM [16] and VLPart [17]. Also, we compare various positional prompting strategies combined with SAM [6]. We will show the efficiency of acquiring patch annotations and their suitability for Guided SAM. It is possible to learn a good segmentation model for a part from only 16 to 64 images, which outperforms state of the art (SOTA) models, while requiring only 5 clicks per image on average.

2 Related Work

For part segmentation, vision-language models have been proposed recently, which can be prompted with a textual description of the part. VLPart [17] trains the model on the part-, object- and image-level to align language and image. An object is parsed by dense semantic correspondence. This approach benefits from various data sources and foundation models, as demonstrated in experiments where the model was applied to unseen object-part combinations (open-vocabulary). OV-PARTS [18] modifies and tailors CLIP [13] for part segmentation. An object mask prompt is proposed to enable the model to take the context into account. To attend more to the parts than whole objects or scenes, a compositional prompt was proposed to reshift attention [16]. Since OV-PARTS and VLPart are both designed to perform open-vocabulary part segmentation, we only consider VLPart in our experiments.

Grounded SAM [16] combines two powerful models: Grounding DINO [7] and SAM [6]. Grounding DINO localizes boxes in the image based on a textual description. Each box contains a prediction where the target may be, in the form of a initial mask. This box is represented by an embedding pair of the top-left corner and the bottom-right corner that serve as positional prompts. These prompts are provided to SAM [6], which segments the target.

Grounded SAM [16] inspired us to look more deeply into the positional prompts themselves. Rather than using the box representations as input for SAM, we aim for a regional prediction that is centered around the part already, to acquire a small but tailored sense of context. Moreover, we want to predict the positional prompts more precisely. For that purpose, we take inspiration from OV-PARTS [18] and Grounded SAM [16], by following their strategy to incorporate some image context around the part. Instead of an implicit context via multiscale annotation (VLPart [17]), we follow OV-PARTS and Grounded SAM by providing an explicit context in the form of a mask or box. Our ROI approach differs because the ROI is more centered around the part, instead of the full object.

3 Guided SAM

Our method segments parts of objects, such as the light of a car, see Fig. 1 (a). Two state of the art methods, Grounding SAM and VLPart, fail on this task, leading to respectively false positives in Fig. 1 (b) and false negatives in Fig. 1 (c). Our objective is to train a capable part segmentation model, while requiring a small amount and labour-efficient type of human annotations. For label-efficiency, we leverage a model M that has strong performance on segmentation already: SAM [6]. This model cannot be applied directly to an image in order to segment a specific part. It requires a spatial cue, provided as a positional prompt $P_{(x,\,y)}$ (a pixel location). Our rationale is to *learn* the spatial cue for a part, in order to guide SAM towards regions in the image where the part is located, P_{ROI}. Our guidance model $\mathcal{G}$ takes an image I and a part C and produces a set of tuples:

$$\mathcal{G}(I \,|\, C) \rightarrow \{(P^i_{ROI},\, P^i_{(x,\,y)})\}_{i \in 1:N} \tag{1}$$

Here, P^i_{ROI} serves as a region of interest (ROI) that conditions where SAM is applied. For each P^i_{ROI}, $P^i_{(x,\,y)}$ serves as the positional prompt for SAM to segment the part. The guidance model $\mathcal{G}$ involves a learner $\mathcal{L}$ that classifies whether an image patch p^j contains the part C: $\mathcal{L}(I \,|\, p^j) \rightarrow c$, where c is the confidence for the part class. Classifying patches is a much simpler learning task than predicting pixel-precise segments. Moreover, the learning requires a simpler form of annotation, i.e., a binary label for the patch if it contains the part or not. Our hypothesis is that such a patch classifier can be learned with a small amount of labels that are simple to annotate. Figure 1 (d) shows the classified patches that are likely to contain the part. A ROI P^i_{ROI} is generated by grouping the classified patches. The positional prompt $P^i_{(x,\,y)}$ for each ROI P^i_{ROI} is inferred from its constituent patches and their respective confidences. Figure 1 (e) shows $(P^i_{ROI}, P^i_{(x,\,y)})$ that was inferred from the classified patches in Fig. 1 (d).

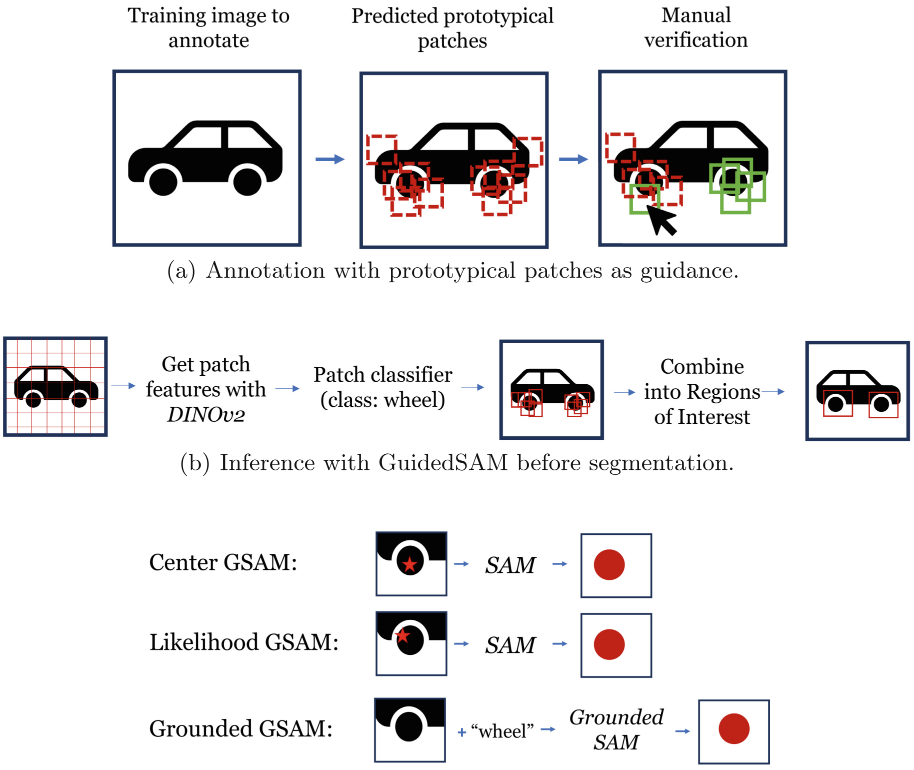
For the segmentation of an object part, both the ROI P^i_{ROI} and the positional prompt $P^i_{(x,\,y)}$ are used. SAM is conditioned on P^i_{ROI}, by passing only the respective image contents. This avoids false positives at irrelevant image regions. SAM is also conditioned on $P^i_{(x,\,y)}$, in order to give it a good starting point for segmentation. Figure 1 (f) shows the part segmentation. Our method enables to use SAM for part segmentation after providing a few labeled patches.

The flow diagram of GuidedSAM is illustrated in Fig. 2. The annotation of patches is shown in Fig. 2 (a) and will be further explained in Sect. 3.1. The inference steps before the segmentation are shown in Fig. 2 (b) and will be covered in Sect. 3.2. Finally, Fig. 2 (c) shows the guided segmentation with multiple model variants that will be explained in Sects. 3.3 and 3.4.

3.1 Prototypical Patches

Our learner $\mathcal{L}$ requires a set of binary labels for respective patches whether they contain the part of interest: $\mathcal{D} = \{(z_i, l_i)\}_{i=1}^{M}$ with M samples, each consisting of a patch z_i and a label $l_i \in [0, 1]$ indicating presence of the part. To arrive at $\mathcal{D}$, the problem is that patches containing parts have a low prevalence, considering that the parts are typically small. Drawing a random selection of patches for annotation, is not efficient. Instead, we select patches that have a larger probability of containing the part. We group similar patches by means of prototypical patches [9]. The prototypes do not have a name, neither are they necessarily related to the part of interest. To relate the prototypes to the part, we match each prototype to the part name, using the visual-textual similarity measure of CLIP [13]. Each prototype is assigned a score for the part of interest. Figure 3 shows examples for various car parts, illustrating that the prototypes group together patches that relate to the respective parts.

For a specific part, the prototypes are ranked by descending CLIP score. Each prototype is verified by a human annotator. For illustration purposes, we indicate this for an example image for the part ‘wheel’ in Fig. 2 (a). This involves one affirmative click if all patches of the prototype contain the part. Similarly, one negative click is required when none of the patches contain the part. More clicks are needed when most patches contain the part, by negating the fewer patches that do not contain it, or vice versa. This procedure yields (z_i, l_i) that constitute $\mathcal{D}$.


Training image to annotate
Predicted prototypical patches
Manual verification
Get patch features with *DINOv2*
Patch classifier (class: wheel)
Combine into Regions of Interest
Center GSAM:
SAM
Likelihood GSAM:
SAM
Grounded GSAM:
+ "wheel"
Grounded SAM

(a) Annotation with prototypical patches as guidance.

(b) Inference with GuidedSAM before segmentation.

(c) Segmentation using different model variants.

Fig. 2. Various elements of the pipeline for GuidedSAM, showing the efficient annotation process in (a), the inference until the segmentation in (b) and the segmentation of the regions of interest with different model variants in (c).

3.2 Guidance Classifier

Given $\mathcal{D} = \{(z_i, l_i)\}_{i=1}^{M}$, the classifier $\mathcal{L}$ is learned, which predicts for a test patch z_j the probability that it contains the part. The patch z_j is represented as a feature vector by a model $\phi(\cdot)$: $z_j^{\phi} = \phi(z_j)$. For $\phi(\cdot)$ we consider DINOv2 [10] which has proven to be a robust feature extractor. $\mathcal{L}_p$ is an SVM [2] with a radial basis function as the kernel. The parameters are learned from train samples $\{(z_i^{\phi}, l_i)\}$. During inference, the trained classifiers are used to predict a rough location of the parts of interest in the test image. This process is illustrated on the left side of Fig. 2 (b).

3.3 Guided Segmentation

A ROI P_{ROI}^{i} is generated by grouping the predicted patches $\{p_j\}$ that are likely to contain the part: $\{\mathcal{L}(I \,|\, p_j) > c_t\}$, where c_t is a threshold on the confidence

c. An example is provided in Fig. 1 (d). The grouping is based on the patches from $\{p_j\}$ that overlap: $\{p_k\}_{IoU>0}$, where $\{p_k\} \subset \{p_j\}$. The ROI P^i_{ROI} is the combination of the minimum and maximum coordinates of the patches in $\{p_k\}$. This is also shown in Fig. 2 (b), where the individual patches are combined into larger ROIs that take into account more context. The positional prompt $P^i_{(x,y)}$ for each ROI P^i_{ROI} is inferred from its constituent patches p_k and their respective confidences c_k by taking the center coordinates of the patch with the highest confidence. Figure 1 (e) shows $(P^i_{ROI}, P^i_{(x,y)})$. Guided SAM is the conditioning of SAM on $(P^i_{ROI}, P^i_{(x,y)})$. An example result is shown in Fig. 1 (f).

3.4 Model Variants

Besides the version of Guided SAM described in Sect. 3.3, we also consider other variants of our conditioning. We make a distinction between applying the segmentation on the ROI P^i_{ROI} (i.e. the combination of the classified patches) or taking the individual patches p_k as ROIs. This is also depicted in Fig. 1 (e), where the bounding box with the dashed line stands for a ROI of combined patches and the smaller box represents an individual patch. For these ROI types there are various options to segment or prompt. Firstly we can replace the segmentation model SAM by Grounded SAM [16] and apply it to the ROI: we coin this model Grounded Guided SAM (GGSAM). This version takes a textual prompt instead of the positional prompt $P^i_{(x,y)}$. Secondly, we can infer the positional prompt $P^i_{(x,y)}$ from the center coordinates of the ROI, which we coin Center Guided SAM (CGSAM). The version that was described before, where the positional prompt is inferred from the center of the patch with the maximum confidence in P^i_{ROI}, is coined Likelihood Guided SAM (LGSAM). This method can only be applied to P^i_{ROI}, since the other ROI is just a single patch. These model variants are illustrated in Fig. 2 (c) on the combined ROIs.

3.5 Computational Load

The computational steps for model inference are shown in Fig. 2 (b). These computations are required on top of the original SAM. We apply an efficient DINOv2 [10] variant to compute the patch features, i.e. ViT-B, which has only 86M parameters. For each part class, the same DINOv2 features are re-used, with a class-specific part classifier. This classifier is an SVM, which involves negligible computations compared to SAM. SAM has 94.7M parameters, comparable to DINOv2 ViT-B, so the computation time of our Guided SAM will be approximately doubled by the classifier guidance. If computational efficiency is essential, faster alternatives are available, e.g. [19], which has a faster backbone. Currently our method applies SAM to every region of interest that is proposed by the part classifiers. This can be implemented more efficiently by re-using its feature maps and only re-running SAM's efficient head on the various regions of interest.

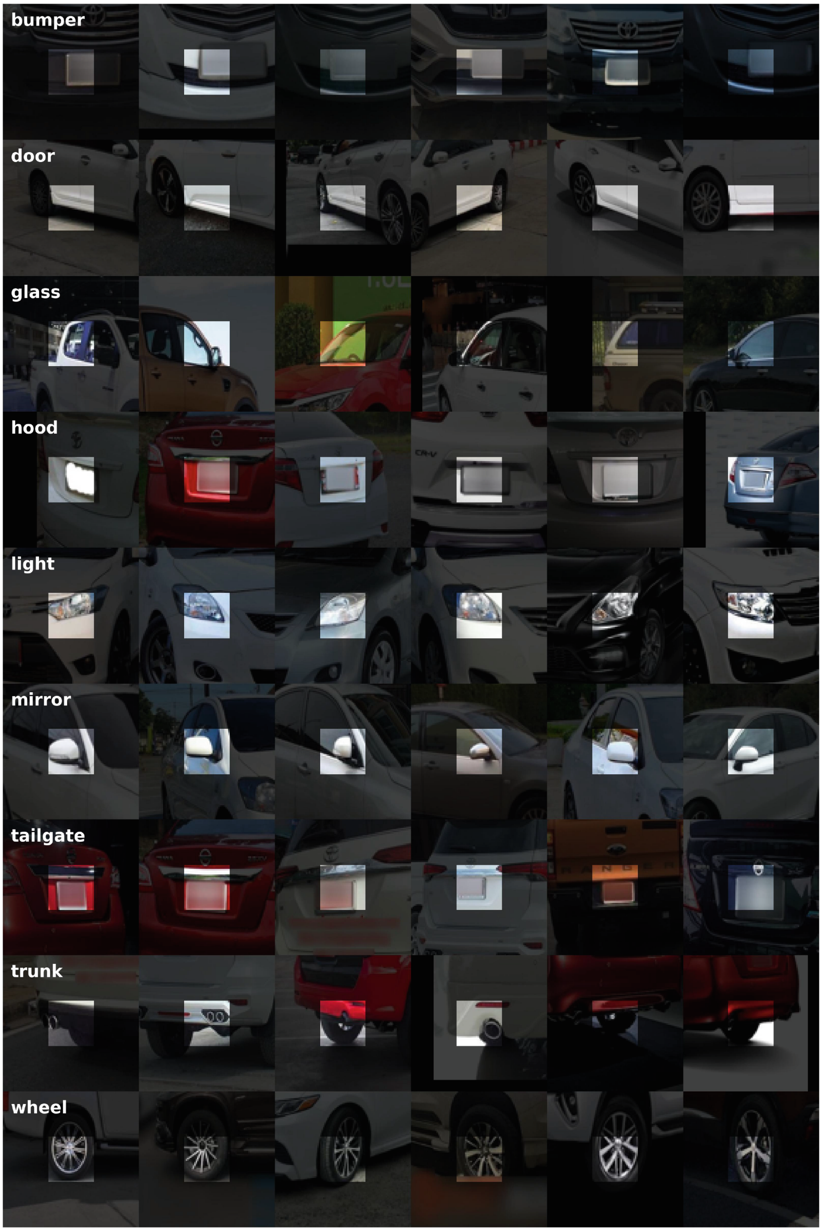
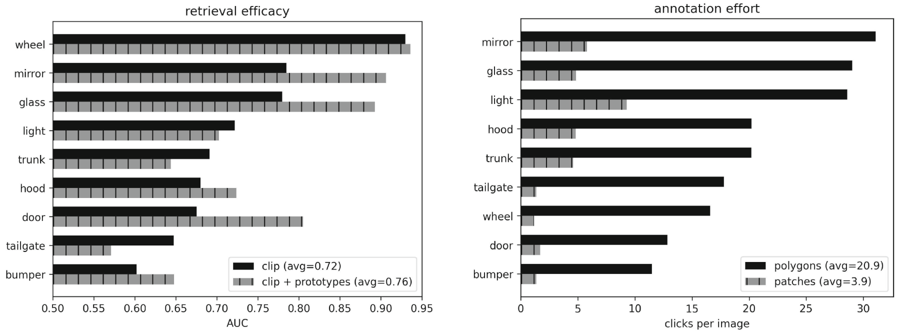


Fig. 3. Prototypical patches group together similar object parts, which facilitates the annotation.

4 Experiments

4.1 Setup

For evaluation we consider the Car Parts Segmentation dataset [12], because of its large inter-class and intra-class variations. The part classes have very different sizes relative to the object. The same part can have different appearances, e.g. forms, sizes and colors. The dataset contains 400 images with annotated segmentation masks of 18 part classes. We merged the different sides (front vs. back, left vs. right) to one part class. There are a total of 9 part classes: bumper, glass, door, light, hood, mirror, tailgate, trunk, and wheel. For language-guided methods (i.e. VLPart and Grounded SAM) we made slight variations to these class names that better reflect the nature of the classes (e.g. replacing glass for window). As a metric, we consider the IoU for each part. The training efficiency is established by increasing the number of training images from 1, 2, 4, 8, ..., 64.



(a) Retrieval efficacy.

(b) Annotation efficiency.

Fig. 4. Prototypical patches are helpful to find the parts (a). Annotating patches is on average >5x more efficient (b).

4.2 Patch Selection

For finding the patches that contain the part, we evaluate the merit of the prototypical patches. To that end, we compare the retrieval efficacy of CLIP with and without the prototypes. Figure 4(a) shows that for most part classes, there is an advantage to consider the prototypical patches. This means that it is helpful to consider the average CLIP score for each prototype before ranking.

To evaluate the annotation efficiency, we compare the amount of manual clicks that are necessary for conventional annotation of polygons to create pixel masks, and for our patch-based annotations using the prototypes. Figure 4(b) shows the annotation efforts for both strategies for the various part classes. On average, annotating patches with our prototype strategy is more than 5 times more efficient. The speedup is most prominent for tailgate and wheel. For conventional annotation, bumper involves the least clicks, because it has a simple shape. Even compared to bumper, all parts are annotated with fewer clicks per image when using the patch-based strategy.

4.3 Guidance Classifier

As shown in Table 1, the learned patch classifier $\mathcal{L}_p$ performs classification of object parts very accurately. For 'door' the performance is the highest: AUC = 0.994 when using 64 training images (in following subsections we will experiment with fewer training images). 'Wheel' also has a very high performance: AUC = 0.990, possibly because of its distinct visual features. For 'trunk' the performance is the lowest, but still very high: AUC = 0.977. Trunk does not have a distinctive boundary and the part is often a flat surface without much texture or distinctive visual features. Light is also somewhat harder to classify (AUC = 0.979). It is a very small part and shows a lot of intra-class variation, such as different shape, size, color (depending on whether it is on or off).

Table 1. The patch classifier performs very accurate classification of object parts (average AUC≈0.985).

Class	door	wheel	mirror	hood	glass	tailgate	bumper	light	trunk
AUC	0.994	0.990	0.989	0.988	0.988	0.987	0.986	0.979	0.977

4.4 Comparison to SOTA

We compare against two methods: VLPart[1] [17] and Grounded SAM [16]. Our method is trained on 64 images. In following subsections we evaluate the impact of having fewer training images. Table 2 reveals that Guided SAM outperforms VLPart and Grounded SAM for most object parts. Overall it is the best performer, on average IoU = 0.493 compared to 0.370 (VLPart) and 0.124 (Grounded SAM). VLPart does perform best at larger or common parts, such as wheel, door and mirror. It is surprising that VLPart does not perform better at other parts, given that it was trained on datasets that include all parts from Table 2, i.e. LVIS [3] and PACO [14]. Grounded SAM performs much worse across the board, probably because it is optimized for objects and not for object parts, although on some parts it performs somewhat better: e.g. wheel and bumper. These are larger parts or parts with clear boundaries. Some parts have a very low performance for both VLPart and Grounded SAM: light, tailgate, and trunk, AUC≈0.04. For these parts, Guided SAM performs much better: AUC≈0.35.

[1] For VLPart we use a confidence threshold of 0.5. For the results of VLPart with varying confidence thresholds, see Supplementary Material.

Table 2. Performance of VLPart [17], Grounded SAM [16], and Guided SAM on the Car Parts dataset in terms of IoU. Bold numbers indicate the best performance per part for the three methods. Guided SAM outperforms VLPart and Grounded SAM for most object parts.

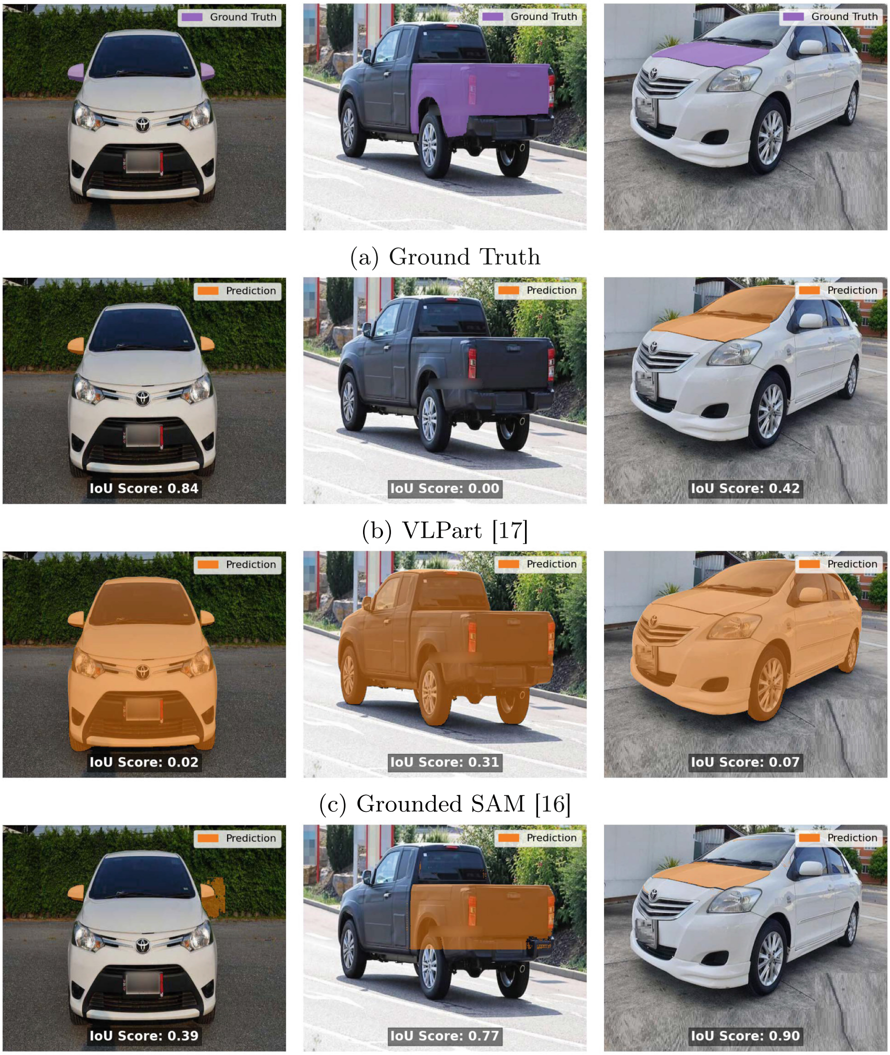
	VLPart	Grounded SAM	Guided SAM (ours)
wheel	**0.800**	0.305	0.683
glass	0.621	0.089	**0.638**
door	**0.736**	0.202	0.635
bumper	0.027	0.299	**0.605**
hood	0.440	0.089	**0.553**
light	0.000	0.041	**0.377**
tailgate	0.000	0.035	**0.370**
trunk	0.006	0.048	**0.314**
mirror	**0.696**	0.009	0.259
average	0.370	0.124	**0.493**

Predictions of the tested models are illustrated for three examples, see Fig. 5. The top row indicates the ground truth, where the other rows show the predicted part segments. VLPart (b) is sometimes very impressive (left), while at other times it misses the part completely (middle), or over-segments it (right). Grounded SAM (c) typically segments the full objects rather than the part. Guided SAM provides a balance, often segmenting the part well, while sometimes over-segmenting or segmenting the background rather than the part.

4.5 Evaluating Model Variants

We evaluate the model variants from Sect. 3.4. As a short recap, we have two main divisions: taking the P^i_{ROI} as the ROI, or its consituent patches $\{p_k\}$ as individual ROIs. This is the top row in Table 3. For each ROI type, there are various options to segment or prompt: Grounded Guided SAM (GGSAM) which uses Grounded SAM as the segmenter, Center Guided SAM (CGSAM) which uses the ROI center as the positional prompt, and Likelihood Guided SAM (LGSAM) which uses the most likely location (i.e. the center of the patch with the highest confidence) as the positional prompt. To establish the effect of the segmentation methods, we also compare with assigning the full patch as the segment, i.e. no segmentation. We refer to this variant as Naive.

Table 3 presents the IoU scores per part for the model variants. ROI guidance is more effective than patch guidance, in most cases. The exceptions are tailgate and trunk, but for these parts ROI guidance performs similarly. Using a positional prompt based on the likelihood (LGSAM) is best on average. There is no single model variant that performs best for all parts. CGSAM performs best on light, hood and mirror. GGSAM appears to perform well on larger car parts that have a distinct boundary, such as bumper, door and wheel. Interestingly,



(a) Ground Truth

(b) VLPart [17]

(c) Grounded SAM [16]

(d) Guided SAM (ours)

Fig. 5. VLPart either segments the object part very well or misses it completely, whereas Grounded SAM typically over-segments severely and often segments the full object. Guided SAM provides a balance, often segmenting the part well, while sometimes over-segmenting.

the performance of GGSAM (i.e. Grounded SAM as segmenter) is much better than applying Grounded SAM on the full image, i.e. without our guidance (Table 2). We conclude that our guidance is also helpful for an existing model.

Table 3. Performance for Region-of-interest (ROI) and Patch guidance for Grounded Guided SAM (GGSAM), Center Guided SAM (CGSAM), and Likelihood Guided SAM (LGSAM) in terms of IoU. Bold numbers indicate the best performance per part for all model variants. ROI guidance is more effective than patch guidance, where segmentation based on likelihood (LGSAM) is best on average.

	Region-of-interest			Patches		
	GGSAM	CGSAM	LGSAM	Naive	GGSAM	CGSAM
bumper	**0.605**	0.319	0.423	0.487	0.560	0.550
glass	0.317	0.626	**0.638**	0.354	0.458	0.480
door	**0.635**	0.447	0.399	0.440	0.508	0.582
light	0.173	**0.377**	0.371	0.170	0.211	0.217
hood	0.395	**0.553**	0.505	0.362	0.406	0.478
mirror	0.063	**0.259**	0.206	0.102	0.113	0.148
tailgate	0.325	0.165	0.370	0.318	0.348	**0.389**
trunk	0.281	0.173	0.314	0.264	0.293	**0.338**
wheel	**0.683**	0.369	0.513	0.246	0.440	0.329
average	0.386	0.365	**0.415**	0.305	0.371	0.390

4.6 Model Selection

There is no single model variant that performs best for all parts (see Table 3). Therefore, we explore model fusion. When selecting the best scores per part out of the three ROI-based methods we get an IoU of 0.493, which is a great improvement over the best model variant (LGGAM with 0.415). We want to understand how many images are needed to decide properly about this model selection. The upper bound is the best-case scenario, established from having seen the full set. The lower bound is worst-case model selection for each part. Now, we are interested in the performance of model fusion when selecting a model variant for each part, after seeing 1, 2, 4, ..., 64 random images. For each amount of images, the experiment is repeated 10 times, because it involves random draws of the images. The increasing performance is shown in Fig. 6. We observe that the average IoU starts way above the lower bound, indicating that just one image is already an indication of which model is most suitable for respective parts. After having seen a few images, e.g. 4 or 8, it is already possible to determine an effective selection of models to acquire better performance by fusion. With 32–64 images, the performance is close to the upper bound.

4.7 Label Efficiency

We hypothesize that the performance of Guided SAM largely depends on the accuracy of the guidance classifier. This classifier is trained with 1, 2, 4, ..., 64 images. We explore how many training images are needed for effective guidance. At various amounts of training images, we evaluate the model variants (Sect. 4.5)

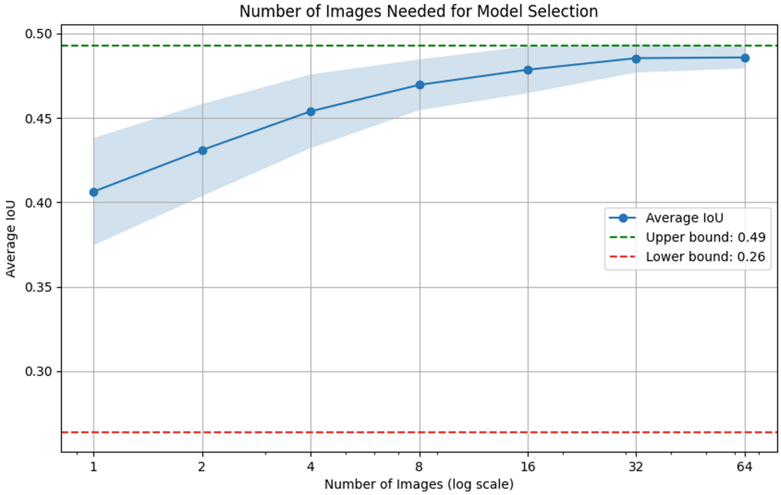


Fig. 6. After having seen a few images, e.g. 4 or 8, it is already possible to determine an effective selection of models to acquire better performance by fusion.

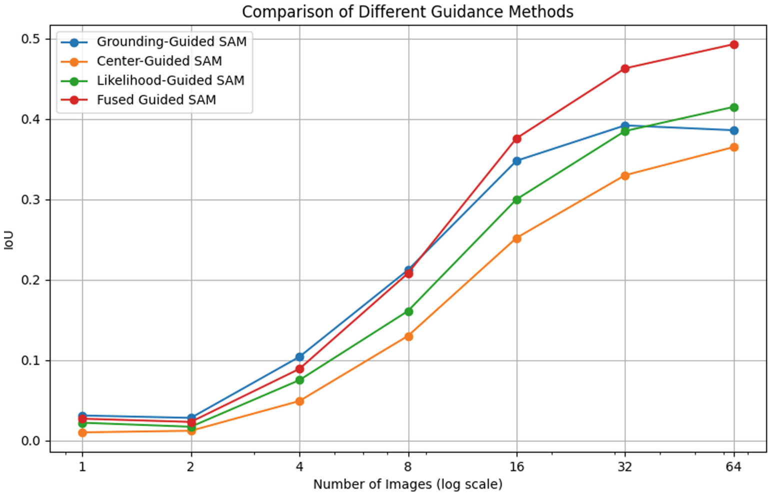


Fig. 7. Learning efficiency of our model variants and the fused model. With 64 images, a performance of IoU≈0.49 is achieved. With 32 or only 16 images, performance drops with respectively only 0.04 or 0.12.

and the fused model (Sect. 4.6), which combines the best performing model variants per part. Figure 7 shows the learning efficiency. The fused model is the best performer on average. Our guidance with Grounded SAM, i.e. Grounded Guided SAM (GGSAM), is the best performer at very low number of images. Probably this is because the confidences of that model are a useful source to filter out wrong segmentations. With more than 8 training images, the fused model has a better performance, especially with 16, 32 or 64 images. With more training images, the guidance becomes better, hence all model variants become better. As a consequence, merit can be taken that the best variant is different across parts (Table 3). With 64 images, a performance of IoU≈0.49 is achieved. With 32 or only 16 images, object parts can be segmented reasonably well: performance drops with respectively only 0.04 or 0.12.

5 Conclusion

In this paper, we proposed a novel method for guiding segmentation models to accurately identify object parts. Our approach leverages regions of interest (ROIs) composed of patches predicted by a learnt classifier to identify specific parts of the object and indicate a positional prompt as starting point for part segmentation. It can be used as a guidance for advanced segmentation models such as (Grounded) SAM. We evaluated our method using the Car Parts dataset and demonstrated that it achieves good performance, even with a limited number of labeled patches. This approach significantly reduces the manual effort required for annotation, as it relies on labeling patches rather than creating full segmentation masks. The patch annotations must be centered around the object parts to ensure that the SAM positional prompts are correctly placed. Misalignment could lead the model to segment the background instead of the intended object parts. For future work, we plan to explore techniques to automatically refine patch placement to enhance segmentation accuracy further. Additionally, we aim to extend our method to other datasets and object categories to validate its generalizability and robustness across various domains.

References

1. Biederman, I.: Recognition-by-components: a theory of human image understanding. Psychol. Rev. **94**(2), 115 (1987)
2. Cortes, C., Vapnik, V.: Support-vector networks. Mach. Learn. **20**, 273–297 (1995)
3. Gupta, A., Dollar, P., Girshick, R.: LVIS: a dataset for large vocabulary instance segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 5356–5364 (2019)
4. Jain, A.K., Hoffman, R.: Evidence-based recognition of 3-D objects. IEEE Trans. Pattern Anal. Mach. Intell. **10**(6), 783–802 (1988)
5. Jia, M., et al.: Fashionpedia: ontology, segmentation, and an attribute localization dataset. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J.-M. (eds.) ECCV 2020. LNCS, vol. 12346, pp. 316–332. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-58452-8_19
6. Kirillov, A., et al.: Segment anything. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 4015–4026 (2023)
7. Liu, S., et al.: Grounding DINO: marrying DINO with grounded pre-training for open-set object detection. arXiv preprint arXiv:2303.05499 (2023)

8. Myers, A., Teo, C.L., Fermüller, C., Aloimonos, Y.: Affordance detection of tool parts from geometric features. In: 2015 IEEE International Conference on Robotics and Automation (ICRA), pp. 1374–1381. IEEE (2015)
9. Nauta, M., Schlötterer, J., van Keulen, M., Seifert, C.: PIP-Net: patch-based intuitive prototypes for interpretable image classification. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 2744–2753 (2023)
10. Oquab, M., et al.: DINOv2: learning robust visual features without supervision. arXiv preprint arXiv:2304.07193 (2023)
11. Palmer, S.E.: Vision Science: Photons to Phenomenology. MIT press (1999)
12. Pasupa, K., Kittiworapanya, P., Hongngern, N., Woraratpanya, K.: Evaluation of deep learning algorithms for semantic segmentation of car parts. Complex Intell. Syst. 1–13 (2021). https://doi.org/10.1007/s40747-021-00397-8
13. Radford, A., et al.: Learning transferable visual models from natural language supervision. In: International Conference on Machine Learning, pp. 8748–8763. PMLR (2021)
14. Ramanathan, V., et al.: PACO: parts and attributes of common objects. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 7141–7151 (2023)
15. Reddy, N.D., Vo, M., Narasimhan, S.G.: CarFusion: combining point tracking and part detection for dynamic 3D reconstruction of vehicles. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1906–1915 (2018)
16. Ren, T., et al.: Grounded SAM: assembling open-world models for diverse visual tasks. arXiv preprint arXiv:2401.14159 (2024)
17. Sun, P., et al.: Going denser with open-vocabulary part segmentation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 15453–15465 (2023)
18. Wei, M., Yue, X., Zhang, W., Kong, S., Liu, X., Pang, J.: OV-PARTS: towards open-vocabulary part segmentation. In: Advances in Neural Information Processing Systems, vol. 36 (2024)
19. Zhao, X., et al.: Fast segment anything (2023)

Multidimensional Cross-Reconstructed Networks for Few-Shot Fine-Grained Image Classification

Yu Cheng, Bo Li(✉), Penghao Jia, Aoxiang Ning, and Jinhong He

Chongqing University of Technology, Chongqing 400054, China
{cy0707,jph0526,ningax,hejh}@stu.cqut.edu.cn, libo@cqut.edu.cn

Abstract. In recent years, numerous Few-Shot Fine-Grained Image Classification methods have been proposed, primarily focusing on better fine-grained feature extraction. Among them, the feature mapping reconstruction network (FRN) is a prominent approach to solving this problem. Nevertheless, extensive comparative experiments reveal that traditional FRN only utilizes support features from a single channel dimension to reconstruct query features, while neglecting interactions between different dimensions, which leads to inaccurate reconstruction errors. To mitigate this issue, this paper proposes a cross-reconstruction network (CRN), which effectively helps the model learn the features across different dimensions, enhancing its applicability to the few-shot fine-grained classification problems. Additionally, we introduce a multi-scale feature enhancement (MCFE) module for feature information, which works in concert with the cross-reconstruction network to capture feature information of images more effectively and make features more specific. Extensive experiments on a baseline dataset demonstrate the superiority of our approach compared to other state-of-the-art methods.

Keywords: Few-shot Fine-Grained image classification · Few-shot learning · Feature reconstruction · Feature fusion

1 Introduction

In recent years, researchers in the computer vision community have devoted considerable attention to few-shot learning [17] [28], particularly few-shot classification [1,5,8,37]. Among these, few-shot fine-grained image classification has posed significant challenges due to the limited number of labeled samples per category and the high similarity between subcategories. Therefore, models need to the capability to learn and distinguish fine-grained features from a small number of labeled samples.

To address the challenges of few-shot classification, researchers have extensively explored meta-learning-based approaches. Meta-learning strives to acquire meta-level knowledge from base classes and subsequently apply this knowledge to new classes. Existing meta-learning methods can generally be divided into three

A. Antonacopoulos et al. (Eds.): ICPR 2024, LNCS 15329, pp. 307–321, 2025.
https://doi.org/10.1007/978-3-031-78110-0_20

categories, metric-based methods [9,30,31,35,41], optimization-based methods [6,23,24,27,28], and transfer learning-based methods [19] [2].

Recently, some metric-based methods have introduced novel alignment [42] or reconstruction [16] [17] [39] techniques, achieving impressive performance in few-shot fine-grained image classification. Among these, the Feature Mapping Reconstruction Network (FRN) [39], proposed by Wertheimer et al., demonstrates excellent performance in few-shot fine-grained classification by weighting and reconstructing each position of its feature map using the ridge regression formula for the support features of each category, and employing the reconstruction error to compute the metric score. However, our experiments reveal that when employing basic features from the usual embedding module for reconstruction, FRN focuses on reconstructing the query features from single-dimensional support features. This approach neglects interactions between different dimensions, which are crucial for fine-grained image classification, leading to inaccurate reconstruction errors. This limitation represents a key issue in the FRN methodology.

Firstly, we developed the multi-scale feature enhancement (MCFE) Module to address the challenge of model performance being affected by background noise and complex scenes in images. The MCFE module enhances computational efficiency through parallel processing and global information encoding, while suppressing the interference of irrelevant information. Furthermore, traditional FRN only reconstructs features from a single channel dimension of the supporting features, ignoring the interactions between different dimensions. This oversight can result in inaccurate reconstruction of the target. To tackle the issue of inaccurate reconstruction, we propose a new reconstruction approach. The method captures the relationship between the different dimensions of the support feature map and the query feature map, thus providing a more accurate distance metric. Our main contributions can be summarized as follows:

- We propose a new multidimensional cross-reconstructed network (FMCRN) for few-shot fine-grained image classification.
- We propose a new cross reconstruction approach (CRN) that replaces the traditional single reconstruction approach.
- We propose a new feature extraction module (MCFE) to extract important feature information in feature maps, which can help in semantic understanding of images.
- The results on fine-grained image datasets, coarse-grained datasets, and difficult datasets with few-shot classification consistently prove the superiority of our proposed method.

2 Related Work

2.1 Few-Shot Learning

In a broad sense, few-shot learning methods based on deep learning can be mainly classified into three categories. The model-based approach [3,7,22,28] aims to quickly update the parameters using a small number of samples by designing the model structure to directly establish the mapping function between the input X and the predicted value P. Optimization strategy-based approaches [6,23,24,27,28] quickly adapt to new tasks by learning how to adjust model parameters.

In this work, we mainly focus on the third type, i.e., metric-based approaches [18]. Metric-based approaches focus on learning a metric space such that samples from the same category are closer to each other and samples from different categories are further away. BSNet [17] uses two similarity metrics to learn the distinct features of each class as a way to classify them. MatchNetwork [35] classifies query samples by calculating match weights for samples in each category, using an attention mechanism to focus on the feature representations in the support set that are most relevant to the query samples, and then using these weighted feature representations to comprehensively assess the similarity of the query samples to known categories.

Our research is not about creating new metrics; instead, it explores how to extract features from metric learning that are critical for category differentiation, with the goal of improving the accuracy of few-shot learning tasks.

2.2 Feature Alignment in Few-Shot Image Classification

Feature alignment methods aim to improve the learned similarity between images by aligning features in an image to capture the spatial location between objects. This helps the model to better utilize the similarities or differences between samples to improve classification accuracy. Feature alignment methods can be broadly categorized into spatial alignment [9,11,39,40,42] and channel alignment [9,11,14,29].

The cross attention module (CAM) in CAN [9] focuses on the semantic correlation between features. It computes the cross attention graph by calculating the correlation between each pair of classes and the query feature graph to highlight common areas to localize objects. The PARN [40] is its ability to compute the similarity between any two positions in an image feature regardless of their spatial distances in the image, which significantly improves the fast adaptation and classification of new categories in sample less learning scenarios. DSN [29] constructs a dynamic classifier by using subspaces to represent the set of samples in each category. This approach exploits the power of subspaces to capture and model structure in high-dimensional data, thereby improving classification accuracy and generalization. RENet [11] performs autocorrelation transformation through self-correlation representation (SCR), allowing the model to extract structured patterns from each image. It also computes cross-correlation between two image representations through cross-correlation attention (CCA), enabling the model to generate joint attention and focus on semantically relevant content between images.

FRN [39] uses ridge regression to reconstruct the feature map of a query image based on support features, providing closed-form solutions that are computationally efficient. Although FRN attempts to preserve the spatial details of the image, it struggles to effectively reconstruct the image features because it does not fully account for the semantic information

Unlike existing reconstruction-based approaches, our proposed FMCRN introduces a new cross-reconstruction module. This module is based not only on channel reconstruction but also on height reconstruction, width reconstruction, and height-width cross-reconstruction. This design allows for a fuller integration of both basic and local content-rich feature representations, thereby enhancing the network's semantic understanding of images.

2.3 Attention Mechanism

Over time, the attention mechanism has emerged as a pivotal technique in enhancing image classification performance. It enables the model to concentrate on essential information within the image while filtering out irrelevant background, thereby enhancing classification accuracy and efficiency.

Since the introduction of the transformer model [34], traditional channel attention mechanisms, such as (SENets) [10], enhanced feature representation by explicitly modeling interrelationships between channels. Subsequently [20] proposed global attention mechanism (GAM), which aims to enhance the performance of deep neural networks by preserving channel and spatial aspects to enhance cross-dimensional interactions. Shortly after, [25] and others proposed a novel efficient multiscale attention (EMA) module focusing on reducing the computational cost while preserving the information of each channel.

Traditional few-shot fine-grained image classification networks often neglect cross-scale information integration, limiting the model's ability to capture subtle yet critical features. To solve this issue, we developed a new multi-scale feature enhancement module (MCFE), inspired by the research of [25]. The core of the design of this module is to address the problem of accurately capturing those small but critical feature differences in learning situations with only a small number of samples, by first reconstructing them as noise maps, and then by synthesizing feature information at different scales to capture those subtle features that are critical for classification. The combined use of various multi-scale feature enhancements significantly improves the model's sensitivity to imperceptible changes in the image, which in turn achieves higher accuracy and reliability in image classification tasks.

3 Methodology

3.1 Definition of the Problem

The few-shot classification problem usually involves dividing the given dataset $D = \{(x_i, y_i), y_i \in C\}$ into three sub-datasets. These are the training set $D_{train} = \{(x_i, y_i), y_i \in C_{train}\}$, the validation set $D_{val} = \{(x_i, y_i), y_i \in C_{val}\}$, and the test set $D_{test} = \{(x_i, y_i), y_i \in C_{test}\}$, which is the final test to evaluate the model. These three sub-datasets are in principle mutually exclusive, so they contain different image classes, i.e. $D_{train} \cap D_{val} \cap D_{test} = \varnothing$.

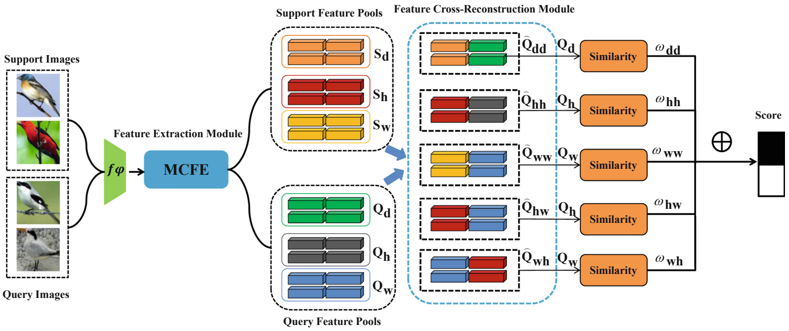


Fig. 1. The raw image is fed to the embedding module f_φ to obtain a basic feature which is subsequently converted by the MCFE module into a richer feature representation. We use orange, yellow and red to denote the three subcategories of the support image, and green, gray and light blue to denote the query image. Then, the cross-reconstruction module cross-reconstructs the two types of query features based on the two types of support features to obtain five reconstruction tasks. ω_{dd}, ω_{hh}, ω_{ww}, ω_{hw} and ω_{wh} denote the corresponding weights used to compute the weighted reconstruction scores of the query images. Finally, the similarity is computed based on the weighted reconstruction errors thus obtaining the metric scores.

N-Way K-shot is a common setup for few-shot, i.e., each training sample consists of N classes randomly sampled from D_{train}, and each class consists of K labeled samples to train the model and L unlabeled samples. The support set and query set are for model learning and validation, respectively. Each class in the support set $S = \{(x_i, y_i)\}_{i=1}^{n} (n = N \times K)$ contains K labeled samples, while each class in the query set $Q = \{(x_q, y_q)\}_{q=1}^{m} (m = N \times L)$ contains L unlabeled samples.

3.2 The Framework of FMCRN

In Fig. 1, we describe the framework of FMCRN. The support set S and the query set Q are input into the embedding module f_φ to extract basic features. The multi-scale feature enhancement module (MCFE) takes the base features as inputs and generates features S_d, S_h, S_w and Q_d, Q_h, Q_w with richer semantic information while reducing the effect of noise. These features are then fed into the cross-reconstruction module, producing five sets of cross-reconstructed query features. The metric distance between these cross-reconstructed query features and their corresponding true query features is computed as the reconstruction error. The weighted sum of these five reconstruction errors is used as a metric score to classify the query image.

3.3 The Multi-scale Feature Enhancement Module (MCFE)

The multi-scale feature enhancement module, as shown in Fig. 2, aims to utilize a convolutional layer to reconstruct the noise mapping and introduces a residual

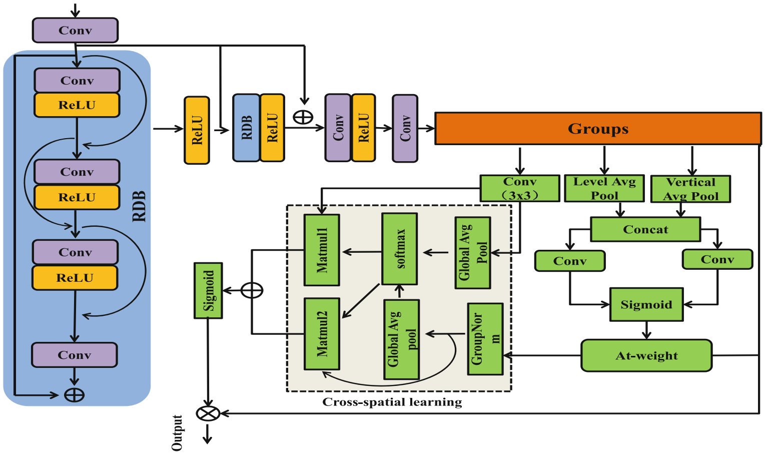


Fig. 2. The structure of MCFE.

learning mechanism that removes redundant features by comparing the enhanced noise-mapped image with the original noisy image. This process enhances the ability to capture features at different scales and improves classification accuracy through fine feature modeling and multi-scale contextual information fusion.

Firstly, we utilize the residual dense architecture [32,43] for feature enhancement, which efficiently identifies and eliminates redundant high-frequency features through a multi-layer structure and residual learning techniques, enhancing the network's noise suppression ability and improving the denoising effect. Subsequently, we pass the enhanced image through a feature extraction module, which effectively extracts and fuses features from different scales and spatial locations, thereby enhancing the richness and accuracy of feature representations for better classification.

Specifically, we first extract image shallow features using a convolutional layer with a 5×5 convolutional kernel size, and further refine the image features using two combinations of RDB [43] and ReLU with a 4-layer structure. Then, we fused the image features learned by two residuals to achieve the enhanced ability of shallow features to memorize deep features. To prevent over-enhancement we pass the fused features through a convolutional layer and ReLU again. We used 1×1 and 3×3 convolutional kernels to extract image features. For the 1×1 convolutional branch, we use global average pooling to encode separately along the horizontal and vertical directions to obtain two parallel feature encoding vectors. These operations can be represented as:

$$Z_H^C = \frac{1}{W}\sum_{j-1}^{W} X_C^{(i,j)}, Z_W^C = \frac{1}{H}\sum_{i-1}^{H} X_C^{(i,j)} \tag{1}$$

Where $X_C^{(i,j)}$ represents the feature at position (i, j) of channel C, and Z_H^C and Z_W^C are the features obtained from pooling in the horizontal and vertical directions, respectively.

For 3 × 3 convolutional branching, we use global average pooling to encode global spatial information as follows:

$$Z_C = \frac{1}{H \times W} \sum_{i-1}^{H} \sum_{j-1}^{W} X_C^{(i,j)} \tag{2}$$

Further, we optimize network efficiency and performance through feature grouping and parallel processing. The feature grouping strategy divides the channel into subgroups, with each subgroup focusing on different semantic features, thereby enhancing feature representation. The parallel processing unit consists of two branches: one captures local features using 1 × 1 convolution and average pooling, while the other captures broader spatial context using 3 × 3 convolution, which accelerates feature extraction and enhances the capture of multi-scale feature. This setup accelerates feature extraction and improves the capture of multi-scale features.

The cross-dimensional interaction mechanism further integrates information from different sub-feature groups. Two spatial attention maps were generated through matrix dot product operations and channel features. The first spatial attention map is obtained by multiplying the output of the parallel processing with the output of the 1 × 1 convolution branch, while the second spatial attention map is obtained by transforming the output of the 1 × 1 convolution branch into the corresponding dimensional shape, retaining precise spatial location information.

Ultimately, the output feature maps within each group are derived by aggregating the two spatial attention weights and applying a specific function. This multi-scale contextual fusion strategy enables the network to generate pixel-level fine attention for high-level feature maps more efficiently, thereby improving feature representation quality. These operations can be expressed as:

$$y_C^{(i,j)} = \sigma(\sum_{k=1}^{C} (Z_K \bullet \omega_{KC}^{(i,j)})) \tag{3}$$

Where σ is the Sigmoid activation function, Z_K is the global spatial information of the Kth channel obtained by global average pooling, and $\omega_{KC}^{(i,j)}$ is the learned attention weights.

3.4 Cross-Reconstruction of Features Based on Ridge Regression(CRN)

The FRN tries to find the matrix $W \in R^{kr \times d}$ to reconstruct Q as a weighted sum of rows in S. In essence the reconstruction of Q depends on the vector of S in the channel dimension. Yet, after our extensive experiments, we have learned that the single-channel dimension-based reconstruction approach ignores the relationship between different dimensions, which leads to inaccurate reconstruction errors.

In order to solve this kind of problem, we design a new feature cross-reconstruction module. In this new feature cross-reconstruction task, we follow the ridge regression strategy proposed in Wertheimer et al. [39] for all reconstruction tasks. The support features $S \in R^{kT \times d}$ and query features $Q \in R^{T \times d}$ (where

$T = h \times w$), three different types of features are newly extracted as S_d, S_h, S_w in the support feature pool, and with the query feature pool Q_d, Q_h, Q_w, and then these features are cross-refactored, resulting in five refactoring tasks.

They are height and width based cross reconstruction, channel based S_d and Q_d reconstruction, height based S_h and Q_h reconstruction and width based S_w and Q_w reconstruction. Which $S_d \in R^{khw \times d}, S_h \in R^{kdw \times h}, S_w \in R^{kdh \times w}$; $Q_d \in R^{hw \times d}, Q_h \in R^{dw \times h}, Q_w \in R^{dh \times w}$.

This cross-reconstruction method is feasible since we get equal height and width in the feature graph, and we reconstruct the query feature Q_w in the width dimension by the support feature S_h in the height dimension and the support feature S_w in the width dimension to reconstruct the query feature Q_h in the height dimension.

$$W_{hw} S_h \approx Q_w, W_{wh} S_w \approx Q_h \tag{4}$$

where $W_{hw} \in R^{dh \times kdw}$, $W_{wh} \in R^{dw \times kdh}$. The formula suggests that this process involves the intersection of two dimensions, Q_w is eventually reconstructed as a weighted sum of S_h rows, Q_h is eventually reconstructed as a weighted sum of S_w rows.

We finally use self-reconstruction based on height, width and channel. That is, S_h reconstructs Q_h, S_w reconstructs Q_w and S_d reconstructs Q_d. Where $W_h \in R^{dw \times kdw}$,$W_w \in R^{dw \times kdh}$,$W_d \in R^{hw \times khw}$.

$$W_h S_h \approx Q_h, W_w S_w \approx Q_w, W_d S_d \approx Q_d \tag{5}$$

To compute the optimal W_{hw}, W_{wh}, W_h, W_w,W_d, we learn from ridge regression by solving the following least squares problem:

$$W = \underset{W}{argmin} ||Q - WS||^2 + \lambda ||W||^2 \tag{6}$$

W is denoted as W_{hw}, W_{wh}, W_h, W_w,W_d, respectively, and λ is denoted as λ_{hw}, λ_{wh}, λ_h, λ_w,λ_d. Following the rules of the FRN [39], where $|| \cdot ||$ is the Frobenius paradigm number and λ is a constant controlling the bias-variance tradeoff, which we set as a learnable parameter. The target weights W have the following closed form solutions:

$$W = QS^T(SS^T + \lambda I)^{-1} \tag{7}$$

where $I \in R^{kT \times kT}$ is the unit matrix, $W \in \{W_{hw}, W_{wh}, W_h, W_w, W_d\}$, $S \in \{S_{hw}, S_{wh}, S_h, S_w, S_d\}$, $Q \in \{Q_{hw}, Q_{wh}, Q_h, Q_w, Q_d\}$, $\lambda \in \{\lambda_{hw}, \lambda_{wh}, \lambda_h, \lambda_w, \lambda_d\}$ Thus, the reconstructed query feature image can be computed as:

$$\hat{Q}_{hw} = W_{hw} S_{hw}, \hat{Q}_{wh} = W_{wh} S_{wh}, \hat{Q}_h = W_h S_h, \hat{Q}_w = W_w S_w, \hat{Q}_d = W_d S_d \tag{8}$$

We compute the similarity of these reconstructed query features with the original unprocessed query features to obtain the reconstruction error of reconstructed query features from the support features for each dimension.

$$E_{hw} =< Q, \hat{Q}_{hw} >= \frac{||Q - \hat{Q}_{hw}||^2}{dh}, E_{wh} =< Q, \hat{Q}_{wh} >= \frac{||Q - \hat{Q}_{wh}||^2}{dw} \tag{9}$$

With the above formula, we can get the reconstruction error based on each dimension. Eventually we need to perform a weighted sum of the reconstruction errors obtained from the appealed formula to get the final reconstruction error.

$$E_r =< Q, \hat{Q} >= \omega_{hw}E_{hw} + \omega_{wh}E_{wh} + \omega_h E_h + \omega_w E_w + \omega_d E_d \tag{10}$$

where ω_{hw}, ω_{wh}, ω_h, ω_w, ω_d, are the learnable weights associated with each of the five reconstruction tasks.

Based on the above reconstruction error, we can get the final classification prediction probability. Where c denotes the category, γ is the learnable temperature factor and C denotes the set of supported categories.

$$P(y_q = c|x_q) = \frac{e^{(-\gamma E_r)}}{\sum_{i \in C^{e^{(-\gamma E_r)}}}} \tag{11}$$

4 Experimental Results and Analysis

4.1 Datasets

To evaluate the effectiveness of the proposed FMCRN model, we conduct extensive experiments on four different fine-grained datasets: the CUB [36], Aircraft [21], Dogs [12] and Cars [13].

The CUB dataset contains 200 bird categories with a total of approximately 600 images per category, totaling 11,788 images. We randomly divided them into a training set of 100 classes, a validation set of 50 classes and a test set of 50 classes. In addition, we crop each image to the bounding box of the human annotations following the preprocessing methods of [41,42].

The Aircraft dataset contains 100 aircraft classes and 10,000 images, and we randomly select 50 classes as the training set, 25 classes as the validation set, and 25 classes as the test set. The aircraft in each image are annotated with tight bounding boxes and hierarchical aircraft model labels.

Dogs Dataset contains 20,580 labeled images from 120 different dog breeds, with about 150–200 images per breed. We randomly selected 70 classes to form the training set, 20 classes for the validation set, and 30 classes for the test set.

The Cars dataset contains 16, 185 images of 196 classes of cars. We randomly select 130 classes to form the training set, 17 classes for the validation set, and 49 classes for the test set.

We choose these four fine-grained datasets as a measure of the performance results achieved by FMCRN in each domain. By using this approach we can make our model more convincing in the domain of few-shot fine-grained classification.

4.2 Implementation Details

Given the small size of our training samples, we use the lightweight Conv-4 as the backbone network, which is commonly used in recent classification work and also known as Conv-64F, as shown in Fig. 3. This network contains 4 identical

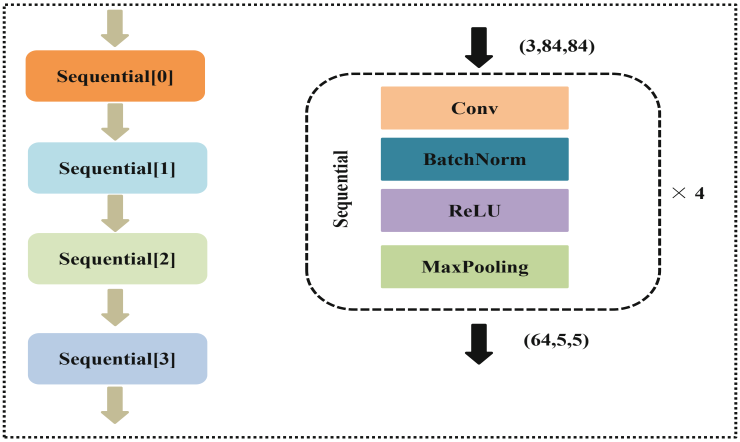


Fig. 3. Structure of the Conv-4 backbone.

convolutional blocks, each consisting of a convolutional layer, a BatchNorm layer, a ReLU activation, and a max-pooling layer of size 2. Thus, when we set the size of the input image to $3 \times 84 \times 84$, the shape of the output feature map is $64 \times 5 \times 5$. During the training phase of our model, we set the epoch to 800 and the initial lr to $1e^{-2}$, weight decay is set to $5e^{-4}$, λ and β are set as learnable parameters ($\lambda \in \{\lambda_{hw}, \lambda_{wh}, \lambda_h, \lambda_w, \lambda_d\}$, $\beta \in \{\beta_{hw}, \beta_{wh}, \beta_h, \beta_w, \beta_d\}$). We validate the model by performing every 20 epochs, thus selecting the model with better performance based on the validation set. These parameters are consistent for all datasets

For all experimental procedures, we use 5-Way 1-shot and 5-Way 5-shot settings for testing and performed 10,000 random samples on the test dataset in order to compute the average classification accuracy with 95% confidence intervals.

4.3 Comparison with State-of-the-Art Technology

In order to verify the applicability of our model in few-shot fine-grained image classification, we conduct experiments on the following four fine-grained image classification datasets. Based on these four fine-grained datasets we reproduce some classical methods for fine-grained image classification. For example (MatchingNet [35], ProtoNet [30], Relation [31], Baseline++ [1]) as well as more recent methods for fine-grained image classification (DSN [29], DN4 [15], DeepEMD [42], RENet [11], the FRN [39], TDM [14], and LCCRN [16]), and as can be seen in Table 1, our method achieves the highest accuracy on these four datasets.

In addition to conducting experiments on the fine-grained dataset, we also use the same training strategy as the fine-grained dataset to conduct experiments on the coarse-grained dataset mini-ImageNet [26] and meta-iNat dataset [33] [38]. The results in Table 2 shows that FMCRN beats FRN and LCCRN in

Table 1. Comparative performance of 5-Way 1-shot and 5-shot on CUB, Aircraft, Dogs and Cars datasets. ♭ denotes our reproduced results.

Model	CUB		Aircraft		Dogs		Cars	
	1-shot	5-shot	1-shot	5-shot	1-shot	5-shot	1-shot	5-shot
MatchingNet [35]	60.06 ± 0.88	74.57 ± 0.73	58.23 ± 0.89	74.90 ± 0.66	46.10 ± 0.78	59.79 ± 0.85	44.73 ± 0.77	64.74 ± 0.72
ProtoNet$^\flat$ [30]	63.79 ± 0.23	82.71 ± 0.16	58.65 ± 0.21	79.62 ± 0.18	46.24 ± 0.21	69.23 ± 0.16	47.16 ± 0.11	69.27 ± 0.19
RelationNet [31]	63.94 ± 0.92	77.87 ± 0.64	61.73 ± 0.98	75.96 ± 0.72	47.35 ± 0.88	66.20v0.74	46.04 ± 0.91	68.52 ± 0.78
DN4 [15]	57.45 ± 0.89	84.41 ± 0.58	68.41 ± 0.91	87.48 ± 0.49	39.08 ± 0.76	69.81 ± 0.69	34.12 ± 0.68	87.47 ± 0.47
DeepEMD [42]	64.08 ± 0.50	80.55 ± 0.71	62.39 ± 0.74	75.46 ± 0.18	46.73 ± 0.49	65.74 ± 0.53	61.60 ± 0.27	72.84 ± 0.37
BSNet(D&C) [17]	62.84 ± 0.95	85.39 ± 0.56	56.51 ± 1.09	70.80 ± 0.81	43.42 ± 0.86	71.90 ± 0.68	40.89 ± 0.77	86.88 ± 0.50
CTX [4]	72.61 ± 0.21	86.23 ± 0.14	67.41 ± 0.10	80.06 ± 0.34	57.86 ± 0.21	73.59 ± 0.16	66.35 ± 0.21	82.25 ± 0.14
FRN$^\flat$ [39]	73.62 ± 0.21	88.24 ± 0.13	53.97 ± 0.21	72.18 ± 0.18	59.91 ± 0.22	78.62 ± 0.15	64.25 ± 0.22	85.46 ± 0.12
TDM$^\flat$ [14]	74.73 ± 0.21	88.95 ± 0.13	69.90 ± 0.23	83.34 ± 0.15	58.77 ± 0.22	77.32 ± 0.15	66.10 ± 0.21	85.89 ± 0.13
LCCRN [16]	75.75 ± 0.22	88.25 ± 0.13	76.24 ± 0.21	87.65 ± 0.12	62.97 ± 0.22	78.57 ± 0.15	71.22 ± 0.21	86.40 ± 0.12
Ours	**76.14 ± 0.21**	**89.31 ± 0.12**	**77.21 ± 0.21**	**88.17 ± 0.11**	**63.69 ± 0.22**	**80.24 ± 0.14**	**74.58 ± 0.21**	**88.95 ± 0.11**

categorizing coarse-grained data. The main reason is that coarse-grained data usually have greater inter-class differences than fine-grained data. Therefore the cross-dimensional acquisition of global features of an image by FMCRN can achieve better classification results than FRN and LCCRN.

Table 2. Performance on mini-ImageNet and meta-iNat.

Model	mini-ImageNet	
	1-shot	5-shot
FRN [39]	53.040.20	70.820.16
LCCRN [16]	53.930.20	70.410.16
Ours	**55.510.20**	**71.790.16**

Model	meta-iNat	
	1-shot	5-shot
Proto [30]	55.340.23	76.460.16
FRN [39]	62.260.23	80.230.16
Ours	**66.960.23**	**83.090.15**

Table 3. Ablation study on the removal of MCFE and CRN in 5-Way 1-shot and 5-shot on CUB, Aircraft, Dogs and Cars datasets.

MCFE	CRN	CUB		Aircraft		Dogs		Cars	
		1-shot	5-shot	1-shot	5-shot	1-shot	5-shot	1-shot	5-shot
×	×	73.08 ± 0.21	88.13 ± 0.13	53.97 ± 0.21	72.18 ± 0.22	59.91 ± 0.22	78.62 ± 0.15	64.25 ± 0.22	85.46 ± 0.13
✓	×	76.08 ± 0.21	88.57 ± 0.12	76.16 ± 0.21	87.42 ± 0.12	63.18 ± 0.22	79.24 ± 0.15	72.72 ± 0.22	88.32 ± 0.11
✓	✓	**76.14 ± 0.23**	**89.31 ± 0.12**	**77.21 ± 0.20**	**88.17 ± 0.18**	**63.69 ± 0.22**	**80.24 ± 0.15**	**74.71 ± 0.21**	**88.95 ± 0.11**

4.4 Ablation Experiment

Based on the fact that in the benchmark method FRN, the query features are reconstructed by means of support feature vectors in the channel dimension and the resulting reconstruction error is also computed in the channel dimension. In contrast, the proposed FMCRN employs channels, heights, lengths, and intersections of height and width to reconstruct the error. In order to fully assess the validity of different combinations of dimensions of reconstruction error, we performed an ablation study. The results given in Table 3 show that in almost all cases, the performance is worst when both modules are removed and the MCFE module has a significant impact on the accuracy improvement. However this module can lead to further improvement in the performance of the model by combining it with our proposed reconstruction method.

Therefore, we argue that relying solely on reconstruction error in the channel dimension overlooks some crucial feature information, and that our proposed reconstruction method effectively addresses this by supplementing the channel dimension reconstruction error.

4.5 Visualization

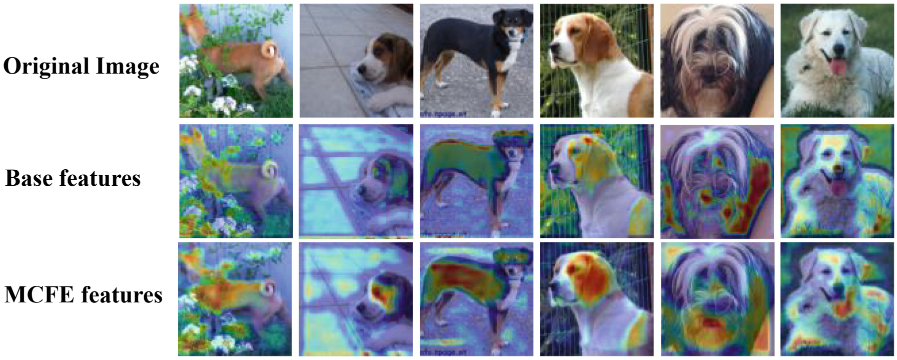


Fig. 4. Visual comparison of base features and MCFE features. Compared to the base module, MCFE extracts richer semantic information and reduces the effect of background.

In Fig. 4, we aim to provide an in-depth comparison of the differences and advantages between the base feature extraction method and our proposed multi-scale feature enhancement (MCFE) technique. The experimental results clearly show that the MCFE technique can effectively reduce the background noise while enhancing the semantic information in the image. By processing images with MCFE, the model captures key information more accurately, leading to improved classification accuracy.

5 Conclusion

In this paper, we propose a multidimensional cross-reconstruction network for few-shot fine-grained image classification. First, we novel a multi-scale feature enhancement (MCFE) module to extract critical information from images. Secondly, our main contribution is a cross-reconstruction module that does not rely on the traditional individual reconstruction errors; We propose a cross-reconstruction module, i.e., a multi-dimensional reconstruction, as well as cross-reconstruction of heights and widths with each other. Compared with existing methods of reconstruction, this method allows us to capture the relationship between different dimensions more efficiently, which is the key to fine-grained learning. Finally, we validate the effectiveness of FMCRN through extensive experiments on four challenging fine-grained datasets, as well as coarse-grained and difficult classification datasets, demonstrating its highly competitive classification performance.

Acknowledgment. This article is supported by the key project of Chongqing basic research (cstc2020jcyjbx0113).

References

1. Chen, W.Y., Liu, Y.C., Kira, Z., Wang, Y.C.F., Huang, J.B.: A closer look at few-shot classification. arXiv preprint arXiv:1904.04232 (2019)
2. Chen, W., Zhang, Z., Wang, W., Wang, L., Wang, Z., Tan, T.: Few-shot learning with unsupervised part discovery and part-aligned similarity. Pattern Recogn. **133**, 108986 (2023)
3. Collier, M., Beel, J.: Implementing neural turing machines. In: Kůrková, V., Manolopoulos, Y., Hammer, B., Iliadis, L., Maglogiannis, I. (eds.) ICANN 2018. LNCS, vol. 11141, pp. 94–104. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-01424-7_10
4. Doersch, C., Gupta, A., Zisserman, A.: CrossTransformers: spatially-aware few-shot transfer. Adv. Neural. Inf. Process. Syst. **33**, 21981–21993 (2020)
5. Fei-Fei, L., Fergus, R., Perona, P.: One-shot learning of object categories. IEEE Trans. Pattern Anal. Mach. Intell. **28**(4), 594–611 (2006)
6. Finn, C., Abbeel, P., Levine, S.: Model-agnostic meta-learning for fast adaptation of deep networks. In: International Conference on Machine Learning, pp. 1126–1135. PMLR (2017)
7. Graves, A., Wayne, G., Danihelka, I.: Neural turing machines. arXiv preprint arXiv:1410.5401 (2014)
8. Guo, Y., Du, R., Li, X., Xie, J., Ma, Z., Dong, Y.: Learning calibrated class centers for few-shot classification by pair-wise similarity. IEEE Trans. Image Process. **31**, 4543–4555 (2022)
9. Hou, R., Chang, H., Ma, B., Shan, S., Chen, X.: Cross attention network for few-shot classification. In: Advances in Neural Information Processing Systems, vol. 32 (2019)
10. Hu, J., Shen, L., Sun, G.: Squeeze-and-excitation networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 7132–7141 (2018)
11. Kang, D., Kwon, H., Min, J., Cho, M.: Relational embedding for few-shot classification. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 8822–8833 (2021)

12. Khosla, A., Jayadevaprakash, N., Yao, B., Li, F.F.: Novel dataset for fine-grained image categorization: Stanford dogs. In: Proceedings of the CVPR Workshop on Fine-Grained Visual Categorization (FGVC), vol. 2. Citeseer (2011)
13. Krause, J., Stark, M., Deng, J., Fei-Fei, L.: 3D object representations for fine-grained categorization. In: Proceedings of the IEEE International Conference on Computer Vision Workshops, pp. 554–561 (2013)
14. Lee, S., Moon, W., Heo, J.P.: Task discrepancy maximization for fine-grained few-shot classification. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 5331–5340 (2022)
15. Li, W., Wang, L., Xu, J., Huo, J., Gao, Y., Luo, J.: Revisiting local descriptor based image-to-class measure for few-shot learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 7260–7268 (2019)
16. Li, X., Song, Q., Wu, J., Zhu, R., Ma, Z., Xue, J.H.: Locally-enriched cross-reconstruction for few-shot fine-grained image classification. IEEE Trans. Circ. Syst. Video Technol. **33**, 7530–7540 (2023)
17. Li, X., Wu, J., Sun, Z., Ma, Z., Cao, J., Xue, J.H.: BSNet: Bi-similarity network for few-shot fine-grained image classification. IEEE Trans. Image Process. **30**, 1318–1331 (2020)
18. Li, Y., et al.: Few-shot fine-grained classification with rotation-invariant feature map complementary reconstruction network. IEEE Trans. Geosci. Remote Sens. **62**, 1–12 (2024)
19. Liu, B., et al.: Negative margin matters: understanding margin in few-shot classification. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J.-M. (eds.) ECCV 2020. LNCS, vol. 12349, pp. 438–455. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-58548-8_26
20. Liu, Y., Shao, Z., Hoffmann, N.: Global attention mechanism: retain information to enhance channel-spatial interactions. arXiv preprint arXiv:2112.05561 (2021)
21. Maji, S., Rahtu, E., Kannala, J., Blaschko, M., Vedaldi, A.: Fine-grained visual classification of aircraft. arXiv preprint arXiv:1306.5151 (2013)
22. Munkhdalai, T., Yu, H.: Meta networks. In: International Conference on Machine Learning, pp. 2554–2563. PMLR (2017)
23. Munkhdalai, T., Yuan, X., Mehri, S., Trischler, A.: Rapid adaptation with conditionally shifted neurons. In: International Conference on Machine Learning, pp. 3664–3673. PMLR (2018)
24. Nichol, A., Achiam, J., Schulman, J.: On first-order meta-learning algorithms. arXiv preprint arXiv:1803.02999 (2018)
25. Ouyang, D., et al.: Efficient multi-scale attention module with cross-spatial learning. In: ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 1–5. IEEE (2023)
26. Ravi, S., Larochelle, H.: Optimization as a model for few-shot learning. In: International Conference on Learning Representations (2016)
27. Rusu, A.A., et zl.: Meta-learning with latent embedding optimization. arXiv preprint arXiv:1807.05960 (2018)
28. Santoro, A., Bartunov, S., Botvinick, M., Wierstra, D., Lillicrap, T.: Meta-learning with memory-augmented neural networks. In: International Conference on Machine Learning, pp. 1842–1850. PMLR (2016)
29. Simon, C., Koniusz, P., Nock, R., Harandi, M.: Adaptive subspaces for few-shot learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 4136–4145 (2020)
30. Snell, J., Swersky, K., Zemel, R.: Prototypical networks for few-shot learning. In: Advances in Neural Information Processing Systems, vol. 30 (2017)

31. Sung, F., Yang, Y., Zhang, L., Xiang, T., Torr, P.H., Hospedales, T.M.: Learning to compare: relation network for few-shot learning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1199–1208 (2018)
32. 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)
33. Van Horn, G., et al.: The iNaturalist species classification and detection dataset. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 8769–8778 (2018)
34. Vaswani, A., et al.: Attention is all you need. In: Advances in Neural Information Processing Systems, vol. 30 (2017)
35. Vinyals, O., Blundell, C., Lillicrap, T., Wierstra, D., et al.: Matching networks for one shot learning. In: Advances in Neural Information Processing Systems, vol. 29 (2016)
36. Wah, C., Branson, S., Welinder, P., Perona, P., Belongie, S.: The Caltech-UCSD birds-200-2011 dataset (2011)
37. Wang, Y., Yao, Q., Kwok, J.T., Ni, L.M.: Generalizing from a few examples: a survey on few-shot learning. ACM Comput. Surv. (CSUR) **53**(3), 1–34 (2020)
38. Wertheimer, D., Hariharan, B.: Few-shot learning with localization in realistic settings. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 6558–6567 (2019)
39. Wertheimer, D., Tang, L., Hariharan, B.: Few-shot classification with feature map reconstruction networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 8012–8021 (2021)
40. Wu, Z., Li, Y., Guo, L., Jia, K.: PARN: position-aware relation networks for few-shot learning. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 6659–6667 (2019)
41. Ye, H.J., Hu, H., Zhan, D.C., Sha, F.: Few-shot learning via embedding adaptation with set-to-set functions. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 8808–8817 (2020)
42. Zhang, C., Cai, Y., Lin, G., Shen, C.: DeepEMD: differentiable earth mover's distance for few-shot learning. IEEE Trans. Pattern Anal. Mach. Intell. **45**(5), 5632–5648 (2022)
43. 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)

Multiplicative RMSprop Using Gradient Normalization for Learning Acceleration

Manos Kirtas[1(✉)], Nikolaos Passalis[1,2], and Anastasios Tefas[1]

[1] Computational Intelligence and Deep Learning Group, Department of Informatics, Faculty of Sciences, Aristotle University of Thessaloniki, Thessaloniki, Greece
{eakirtas,passalis,tefas}@csd.auth.gr

[2] Department of Chemical Engineering, Faculty of Engineering, Aristotle University of Thessaloniki, Thessaloniki, Greece

Abstract. Although deep learning (DL) architectures achieve state-of-the-art performance in a wide range of applications, such as computer vision, the training process remains highly sensitive to hyperparameters, initial weights, and data distributions, making the development of fast and stable optimization methods a challenging task. The Root Mean Square propagation (RMSprop) optimization method has successfully extended the Stochastic Gradient Descent (SGD), using an adaptive learning rate mechanism, establishing its use in the DL community. However, even RMSprop suffers from convergence issues related to the high variance of gradients and learning rates at the initial stage of training. Motivated by the significant contribution of the multiplicative updates in the early development of Machine Learning and recent preliminary results, in this work, we propose a multiplicative update term oriented to RMSprop, significantly improving its performance. More specifically, the proposed term employs normalization to gradients and scales the parameters according to their magnitudes, leading to significant acceleration at the initial stage of training, while resulting in more robust models. Based on the proposed update term, we formulate two novel RMSprop alternatives demonstrating the acceleration and robust capabilities on traditionally used image classification benchmarks as well as to convex and non-convex optimization tasks.

Keywords: RMSprop · Multiplicative updates · Accelerate training · Robust training

1 Introduction

Even nowadays, where Deep Learning (DL) has achieved state-of-the-art performance in computer vision applications, training fast and robust DL models remains a challenging task [1]. Stochastic Gradient Descent (SGD) undoubtedly holds the credential of having a tremendous impact on the early development of DL. In turn, RMSprop, inspired by Adagrad [2], is proposed as an adaptive learning rate alternative to SGD, significantly improving the performance of DL

A. Antonacopoulos et al. (Eds.): ICPR 2024, LNCS 15329, pp. 322–336, 2025.
https://doi.org/10.1007/978-3-031-78110-0_21

models when gradients are sparse or, in general, small. Even though RMSprop, along with Adam, is still missing a rigorous and well established theoretical analysis, is considered as a preferable solution in a wide-range of applications, including image classification, empirically shown state-of-the-art performance.

Although RMSProp improved the rapid decay issue of the learning rate of Adagrad, it still suffers from convergence issues related to the high variance in learning rate during early stages of training [3] or the magnitude of gradients [4], which provides, in turn, a fruitful area for analysis, variations and enhancements [5]. In our work, we focus on the update term of RMSprop, introducing an alternative multiplicative term. More specifically, preliminary results have shown that multiplicative updates leverage advantages over additive updates, due to their properties to involve parameter magnitude during the update, and despite the fact that they have been extensively studied during the early years of machine learning research [6] holding strong theoretical guarantees [7], they have been merely studied in the context of DL. In fact, multiplicative update remains a valuable solution for non-negative matrix factorization, since they physically constrain the sign of parameters, allowing part-based representation, and the authors in [8] highlight the acceleration capabilities that are offered. Such physical restrictions can also be used to preserve the balance between excitatory and inhibitatory synapses of neural networks [9], with recent work trying to exploit such effects to stabilize training [10] or even train the DL architecture with lower precision arithmetic [11].

In this work, we exploit the properties of the multiplicative updates to propose a novel update term, oriented to RMSprop, targeting to accelerate training by overcoming known limitations that are related to gradient and parameter magnitudes. More specifically, the proposed update term normalizes the gradients, providing a simple measurement of how much a single gradient descent step will scale the original parameter, ignoring the magnitude of the gradient and considering the parameter's one, which empirically is shown to make training more robust and faster [4,12]. To fully exploit the potential of multiplicative updates by overcoming the physical sign constraint, we integrate the proposed term into a novel hybrid additive-multiplicative RMSprop approach, showing that both proposed updates can generalize their acceleration and robust behavior, which is observed on simple convex and non-convex optimization tasks on traditionally used image classification benchmarks. The proposed methods offer acceleration during the initial stage of training, while leading to greater robustness to initial distribution models.

To this end, the main contributions of this work are twofold: a) a multiplicative-based update rule that accelerates convergence of the optimization process while making models more robust to initial parameter distributions and b) a hybrid multiplicative-additive RMSprop that overcomes the sign limitation of the multiplicative update term. We provide experimental results on convex and non-convex task, shedding light on the benefits of the proposed updates, while we demonstrate their capabilities on CIFAR image classification benchmarks, employing the proposed optimizers in several scenarios, focusing on the robustness and acceleration capabilities.

The rest of the paper is structured as follows. First, we introduce the proposed methods in Sect. 2. Then, the results of the experimental evaluation are reported in Sect. 3. Finally, the conclusions are discussed in Sect. 4.

2 Proposed Method

Deep neural networks (DNNs) have the ultimate goal of approximating a function f^* using a universal approximator F. More precisely, the input of the networks is indicated as $\boldsymbol{x} \in \mathbb{R}^{N_{i-1}}$, where N_{i-1} represents the number of input features at i-th layer. Each sample in the train data set is labeled with a vector $\boldsymbol{l} = \mathbf{1}_c \in \mathbb{R}^C$, where the c-th element equals to 1 and the other elements are 0 if it is a classification task (C denotes the number of classes) or a continuous vector $\boldsymbol{l} \in \mathbb{R}^C$ if it is a regression task (C denotes the number of regression targets). DNNs approximate f^* by using more than one layer, i.e., $F(\boldsymbol{x};\boldsymbol{\Theta}) = f^{(n)}(\dots(f^{(2)}(f^{(1)}(\boldsymbol{x};\boldsymbol{\theta}^{(1)})\boldsymbol{\theta}^{(2)};)\boldsymbol{\theta}^{(n)}) = \boldsymbol{z}^{(n)}$ and learn the parameters $\boldsymbol{\theta}^{(i)}$ where $0 \leq i \leq n$ with $\boldsymbol{\theta}^{(i)}$ consisting of weights $\boldsymbol{w}^{(i)} \in \mathbb{R}^{N_i \times N_{i-1}}$ and biases $\mathbf{b}^{(i)} \in \mathbb{R}^{N_i}$. For example, the multilayer perceptrons compute the linear output of each layer as:

$$\boldsymbol{z}^{(i)} = f^{(i)}(\boldsymbol{y}^{(i-1)}) = \boldsymbol{w}^{(i)}\boldsymbol{y}^{(i-1)} + \boldsymbol{b}^{(i)} \in \mathbb{R}^{N_i}. \tag{1}$$

The output of the linear part of a neuron is fed to a non-linear function $g(\cdot) : \mathbb{R}^{N_i} \rightarrow \mathbb{R}^{N_i}$, named activation function, to form the final output of the layer:

$$\boldsymbol{y}^{(i)} = g(\boldsymbol{z}^{(i)}) \in \mathbb{R}^{N_i}. \tag{2}$$

Without loss of generality, this can be appropriately generalized to describe convolution neural networks (CNN), with the difference that they apply multidimensional convolutional operations between the features and the kernel parameters. Consequently, training a CNN is achieved by updating its parameters, using the backpropagation algorithm, to minimize an objective $J(\boldsymbol{\theta})$, we use lowercase $\boldsymbol{\theta}$ to simplify the notation. Cross-entropy loss is often used in multi-class classification as objective function, given by:

$$J_t(\boldsymbol{\theta}_t) = -\sum_{c=1}^{C} l_c \log F_c(\boldsymbol{x};\boldsymbol{\theta}_t) \in \mathbb{R}, \tag{3}$$

where t define the training epoch and $F_c(\boldsymbol{x};\boldsymbol{\theta}_t)$ denotes the c-th element of the $F(\boldsymbol{x};\boldsymbol{\theta}_t)$ output.

In the case of RMSprop, the algorithm moves each parameter $\boldsymbol{\theta}_{t-1}$ in the opposite direction of the gradient $\boldsymbol{g_t} = \nabla_\theta J_t(\boldsymbol{\theta}_{t-1})$ adapting the learning rate for each parameter individually using the sequence of gradient estimates. More specifically, the updated parameters is calculates as:

$$\boldsymbol{\theta}_t = \boldsymbol{\theta}_{t-1} - \frac{\eta}{\sqrt{\boldsymbol{E}[\boldsymbol{g}^2]_t + \epsilon}} \in \mathbb{R}^N, \tag{4}$$

where η defines the step size, which by default is set to $\eta = 0.001$ and $\boldsymbol{E}[\boldsymbol{g}^2]_t$ is the moving average of squared gradient:

$$\boldsymbol{E}[\boldsymbol{g}^2]_t = \beta \boldsymbol{E}[\boldsymbol{g}^2]_{t-1} + (1-\beta)\boldsymbol{g}_t^2 \in \mathbb{R}^N, \tag{5}$$

where β is the weighting parameter of moving average and by default is set to $\beta = 0.9$.

However, RMSprop can be sensitive to the initial conditions of parameters; for instance, if the initial gradients are large, the learning rates will be low for the remaining training. Furthermore, the accumulation of squared gradients in $\boldsymbol{E}[\boldsymbol{g}^2]_{t-1}$ can significantly eliminate the learning rate to values close to zero after some training epochs. Therefore, this makes RMSprop sensitive to the choice of learning rate and initialization scheme, leading to difficulties in training DL models when the magnitude of the parameters differs from the initial theoretical hypothesis. To overcome these limitations, we propose a multiplicative update rule oriented to faster convergence and robustness that normalizes and clips gradients. More specifically, the proposed update rule incorporates $\tanh(\cdot) : \mathbb{R} \to (-1, 1)$ function that offers normalization of the gradient and then multiplies it by the parameter. In this way, the proposed update term proportionally scales the parameters considering the magnitude of it, calculated as:

$$\boldsymbol{\theta}_t = \boldsymbol{\theta}_{t-1} - |\boldsymbol{\theta}_{t-1}| \tanh\left(\frac{\eta_{in}}{\sqrt{\boldsymbol{E}[\boldsymbol{g}^2]_t} + \epsilon}\right)\eta_{out} \in \mathbb{R}, \tag{6}$$

where $\eta_{in} \in \mathbb{R}^+$ is the inner and $\eta_{out} \in (0, 1]$ the outer learning rate. Essentially, the inner learning rate allows one to adjust the gradients regarding the working range of the used nonlinearity. The outer learning rate affects the size of the step similarly to the learning rate used in traditionally applied optimization methods and, additionally, defines the upper scaling threshold depending on the parameter, $[-\eta_{out}|\boldsymbol{\theta}_{t-1}|, \eta_{out}|\boldsymbol{\theta}_{t-1}|]$. For a deep learning application, we propose setting by default the learning rates of multiplicative RMSprop as $\eta_{in} = 0.4$ and $\eta_{out} = 0.2$

Motivated by the observation that gradient clipping in a specific setting can accelerate training [13], making DL models more robust [4], the proposed multiplicative update rule naturally normalizes gradients and introduces a threshold proportional to the magnitude of the parameter. More specifically, the proposed multiplicative update rule makes the update term proportional to the parameter by introducing the $\tanh(\cdot) : \mathbb{R} \to (-1, 1)$ function. Intuitively, the multiplier $\tanh(\eta_{in} m_t l_t)$ provides a measurement of how much a single gradient descent step will scale the original parameter. In this way, the divergence issues occurred when the update term $||\eta \nabla_\theta J_t(\boldsymbol{\theta}_t)||$ becomes significantly larger than the weight $||\boldsymbol{\theta}_t||$ can be partially eliminated, where $|| \cdot ||$ denotes the L2 norm.

In fact, as already shown, the ratio of the L2-norm of weights and gradients, $||\boldsymbol{\theta}_t||/||\nabla_\theta J_t(\boldsymbol{\theta}_t)||$, is not only significantly high in the first epochs of training, but also highly different between weights, biases, and layers [14]. As a result, vanishing and exploiting gradient phenomena are prevalent during the initial training stage, making the traditional optimizer highly sensitive to initialization and learning rate [15–17]. Using the proposed method, the update of parameters no longer depends on the magnitude of the gradient, preventing the gradient-weight ratio $||\boldsymbol{\theta}_t||/||(|\boldsymbol{\theta}_t|\tanh\left(1/\boldsymbol{E}[\boldsymbol{g}^2]_t\right))||$ to become significantly large, while it introduces thresholds for maximum increment and decrement that can be easily controlled by the outer learning rate. Thus, we claim that the proposed update term makes training more robust to vanishing and exploiting gradience phenomena, offering acceleration to convergence.

Although the proposed update rule has the ability to preserve the initial sign of the update parameter, making it an excellent choice for training neural networks oriented to interpretability [18] and neuromorphic architectures [19,20], in traditional DL training, there are no such limitations. However, current neural network architectures are initialized by drawing parameters from a normal distribution, $\mathcal{N}(0, \sigma^2)$, with zero mean and variance depending on the size of the layers [21]. Furthermore, by default, DL architectures are overparameterized, allowing optimization methods to converge even in complex loss topologies, stabilize and accelerate training, and, in turn, improve overall performance [22]. At the same time, overparameterization allows different optimization methods to converge at different local minima, leading to equivalent performances [23].

In this work, we exploit overparameterization to apply multiplicative updates that accelerate training, since it allows the proposed multiplicative update rule to converge in a local minimum using the initial sign of the parameters. Where needed, the multiplicative update rule can be combined with the additive one, exploiting in this way the benefits of the multiplicative update rule with the ability of parameters to change sign when additive rule is used. The proposed hybrid rule retains the advantages of multiplicative updates while controlling the contribution of each update, and is given by:

$$\begin{aligned} \boldsymbol{\theta}_t = \boldsymbol{\theta}_{t-1} - \gamma \left[|\boldsymbol{\theta}_{t-1}| \tanh\left(\frac{\eta_{in}}{\sqrt{\boldsymbol{E}[\boldsymbol{g}^2]_t + \epsilon}} \right) \eta_{out} \right] \\ + (1-\gamma)\left(\frac{\eta}{\sqrt{\boldsymbol{E}[\boldsymbol{g}^2]_t + \epsilon}} \right) \in \mathbb{R}^N, \end{aligned} \tag{7}$$

where γ is the weight of the relative contribution of the multiplicative term. By default, we suggest setting $\gamma = 0.5$ and learning rates as $\eta_{in} = 0.01$ and $\eta_{out} = 0.1$. The learning rate of the traditionally additive update term can be set to its default values $\eta = 0.001$

Finally, the hybrid method, except for allowing parameters to change their initial sign when using the multiplicative update rule, also ensures that the parameters will not be stuck at zero, which is a potential consequence of utilizing multiplicative update rules. Although modern DL frameworks, such as PyTorch [24] and Tensorflow [25], do not initialize weights and biases to zero, combining the multiplicative update term with the additive one allows one to overcome such potential limitation, leveraging the advantages of the multiplicative update. However, it should be mentioned that the elimination of synapses could potentially be useful in cases where weight sparsity and/or pruning are required [26], while it can also provide an additional regularization effect, avoiding in this way overfitting during training.

Algorithm 1: RMSprop

```
   Input      : $\eta, \eta_{in}, \eta_{out}$ : stepsizes, $\gamma$: update contribution, $\theta_0$: parameters, $f(\theta)$:
                objective, $\alpha$: alpha, $\lambda$: weight decay,
   Initialize: $u_0 = 0$: square average, $\boldsymbol{b}_0 = 0$: buffer, $g_0^{ave} = 0$
 1 begin
 2     while $t = 1$ to $T$ do
 3         $g_t = \nabla_\theta f_t(\theta_{t-1})$;
 4         if $\lambda \neq 0$ then
 5             $g_t = g_t + \lambda\theta_{t-1}$ ;
 6         $u_t = \alpha u_{t-1} + (1-\alpha) g_t^2$ ;
 7         $\tilde{u_t} = u_t$ ;
 8         // Additive
 9         $\boxed{\theta_t = \theta_{t-1} - \eta g_t / (\sqrt{\tilde{u_t}} + \epsilon) \;;}$
10         // Multiplicative
11         $\boxed{\theta_t = \theta_{t-1} - |\theta_{t-1}| \tanh\big(\eta_{in} g_t / (\sqrt{\tilde{u_t}} + \epsilon)\big)\eta_{out};}$
12         // Hybrid
13         $\boxed{\begin{array}{l} \theta_t = \theta_{t-1} \\ -\gamma\left(|\theta_{t-1}| \tanh\big(\eta_{in} g_t / (\sqrt{\tilde{u_t}} + \epsilon)\big)\eta_{out}\right) \\ -(1-\gamma)\left(\eta g_t / (\sqrt{\tilde{u_t}} + \epsilon)\right) \;; \end{array}}$
14     return $\theta_t$ ;
```

The proposed methods can be easily implemented using the original RMSprop algorithm by changing the update rule, as presented in Algorithm 1. More specifically, in order to integrate the proposed methods, one has to replace the original additive update of RMSprop, given in line 9 (blue color), with either the multiplicative or hybrid, in lines 11 (red color) and line 13 (green color), respectively. In such a way, the leaning capability of the RMSprop algorithm can be easily accelerated, providing a stable training process regardless of the initial distribution.

3 Experimental Results

We experimentally evaluated the proposed framework using two different sets of experiments. First, we demonstrate the effectiveness of the proposed optimization framework in 2-dimensional convex and non-convex tasks, giving us further insights into the optimization process. Then, we apply the proposed optimization method in traditionally used image classification benchmarks employing DNNs.

3.1 Convex and Non-convex Optimization

For a better understanding of the optimization method under the proposed framework, we demonstrate different optimizers in 2-dimensional convex and non-convex tasks. For convex task, we used a second-order polynomial given by:

$$f_1(\boldsymbol{x}) = \beta(x_1 - \alpha)^2 + 10\beta(x_2 - \alpha)^2 \in \mathbb{R}, \tag{8}$$

where $\boldsymbol{x}^* = [\alpha, \alpha]$ is the global minimum of the convex function with subscript i denoting the i-th element of the vector, and β defines the steepness of the function. Additionally, we use the non-convex Rosenbrock benchmark function, given by:

$$f_2(\boldsymbol{x}) = (\alpha - x_1) + \beta(x_2 - x_1^2)^2 \in \mathbb{R}, \tag{9}$$

where the global minimum is at $\boldsymbol{x}^* = [\alpha, \alpha]$ and β defines the steepness of the function. The values of α and β are set to 1 and 20, respectively. The vector parameter $\boldsymbol{x}$ is optimized to solve the equations.

Table 1. Task configurations for tuning and evaluation processes

	Convex 2D		Rosenbrock	
Parameter	Tuning	Evaluation	Tuning	Evaluation
$x_1^{(t=0)}$	50	$\mathcal{N}(50, 5)$	0.5	$\mathcal{N}(0.5, 0.1)$
$x_2^{(t=0)}$	50	$\mathcal{N}(50, 5)$	3.0	$\mathcal{N}(3, 1)$
α	1	$\mathcal{N}(1, 1)$	1	1
β	20	$\mathcal{N}(20, 2)$	60	$\mathcal{N}(60, 6)$
Iterations	100	$\mathcal{N}(100, 10)$	100	$\mathcal{N}(100, 10)$

We evaluate the traditionally used RMSprop and the proposed multiplicative and hybrid updates to find appropriate learning rates for the applied simplified tasks. We tune the hyperparameters of each optimization approach evaluated for the aforementioned tasks with a given configuration. The configuration used for each task is reported in the Table 1 at columns 2 and 4. For hyperparameter tuning, a grid search method is used. More specifically, for the traditionally used additive RMSprop, we perform a search for the learning rate, ranging from $[1e - 6, 5e + 2]$, with a step of 0.5. For the proposed multiplicative update rule, we perform a grid search for inner and outer learning rates, between values $[1e - 1, 5e + 1]$ and $[1e - 4, 1.0]$, using step 0.5, respectively. Finally, for the hybrid variation of the RMSprop, the above learning rate search spaces are applied, while the weight of the relative contribution of the multiplicative term, γ, is kept fixed and equal to 0.5. In all cases, 20 learning rate configurations are evaluated as obtained from the grid search algorithm.

In Table 2, we present the final scores obtained from the tuning process on lines 3 and 4. More specifically, in column 2, we present the performance of traditionally used RMSprop and in columns 3 and 4 the performance for the

Table 2. The table reports in lines 3 and 4 the best scores after 100 iteration steps applying the best hyperparameters as obtained from tuning process in given configurations. At lines 6 and 7, the average test score over 100 randomly drawn task configurations is reported applying the best hyperparameters as obtained from tuning process.

Optimizer	Baseline	Proposed **Multiplicative**	Proposed **Hybrid**
Tuning			
Convex 2D	2.62×10^{-7}	1.09×10^{-7}	$\mathbf{2.42 \times 10^{-9}}$
Rosenbrock	2.22×10^{-1}	8.98×10^{-2}	$\mathbf{2.45 \times 10^{-2}}$
Testing			
Convex 2D	8.26×10^{-6}	2.24×10^{-3}	$\mathbf{2.01 \times 10^{-8}}$
Rosenbrock	3.16×10^{-1}	1.43×10^{-1}	$\mathbf{2.01 \times 10^{-1}}$

proposed multiplicative and hybrid RMSprop alternatives, respectively. For a better comparison, we report the score for each evaluation run, which is computed by dividing the Euclidean distance between the optimized parameters and the global minimum with the Euclidean distance between the initial parameter values and the global minimum. In this way, the zero means that the parameters are on the global minimum, while the one means that the parameters are on the initialization point.

From hyperparameter tuning results, presented in lines 3 and 4, it is highlighted that the proposed hybrid RMSprop performs significantly higher than the traditionally used RMSprop, covering a larger distance from the global minimum during 100 iterations. This confirms that optimal performance is obtained when the multiplicative update rule is combined with the additive one under the proposed hybrid RMSprop. Hybrid RMSprop significantly outperforms both additive and multiplicative update rules, since it exploits the acceleration capabilities of the multiplicative update, while allowing the parameters to change sign. This is demonstrated in Fig. 1 that depicted the Euclidean distance during training. Although the proposed multiplicative update rule results in the worst local minimum than the proposed hybrid RMSprop, it offers a significant performance boost during the first epochs of training. In turn, the hybrid update rule exploits such a boost by combining the proposed multiplicative update term with the additive one, which allows the parameters to change sign, leading to a higher convergence rate and resulting in a better local minimum.

Additionally, we investigate the robustness of the proposed framework, applying the best hyperparameter configurations obtained from the hypeparameter tuning process. More precisely, we draw the initial points, β value (which defines the slope of the function), and the number of iterations of a Gaussian distribution, with a mean equal to the value used in the tuning process and a standard deviation that is proportional to the magnitude of the mean, as reported in Table 1. In addition, in the Convex2D case, the global minimum is drawn from

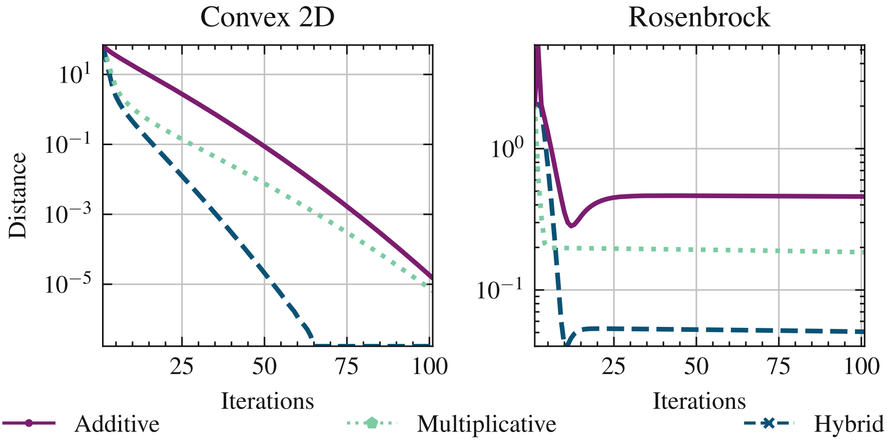


Fig. 1. The figure depicts the Euclidean distance between parameters and the global minimum for Convex 2D and Rosenbrock tasks when applying alternative updates for RMSprop optimizer.

a normal distribution as well. We report the average score over 100 evaluation runs for the three optimizers in Table 2 in rows 6 and 7.

Evaluating the different RMSprop approaches in a randomly drawn configuration shows that the proposed hybrid RMSprop leads to significantly better performance, while the multiplicative alternative suffers from a convergence issue related to the sign constraint. More specifically, in the Convex 2D case, the improvements when the hybrid update rule is used are impressive compared to the traditionally used updates. Similar improvements are also achieved in the Rosenbrock non-convex optimization task. The performance of hybrid updates highlights the generalization ability of the proposed alternative on non-convex task, while by comparing its performance with the one obtained when using the multiplicative RMSprop, it is shown that the hybrid update rule exploits the benefits of multiplicative updates overcoming the restrictions forced by the multiplicative update. In this way, the robustness and generalizability of the proposed hybrid RMSprop are demonstrated, achieving better performance than traditional updates in cases where the initial theoretical hypothesis differs from the actual one.

3.2 Image Classification

To demonstrate the capabilities of the proposed RMSprop alternatives, we evaluate them in traditionally used image classification benchmarks. More specifically, we conducted experiments to evaluate the robustness and convergence of the proposed optimizers in DNNs applying different size architectures to traditionally used image classification benchmarks. First, we present experimental results in the CIFAR10 dataset, which consists of 600,000 32×32 colored images in 10 classes, with 6000 images per class, where 50,000 are training images and 10,000 test images, applying different training configurations and DNN architectures, such as ResNet9, ResNet18 and VGG16 [27,28]. More specifically, 10 random training configurations are used, applying Xavier initialization[1] with different gain values and different number of epochs, highlighting in this way the robustness to the initial distribution when the proposed multiplicative update rule is applied. Gain values are randomly drawn from a gamma distribution, defined as $g \sim Gamma(k = 1, \theta = 2.5)$, while the number of training epochs is drawn from a normal distribution, defined as $T \sim \mathcal{N}(60, 10)$. The default hyperparameter values are used for each different optimizer, while we apply a simple preprocessing to training data involving: a) random crop, b) random horizontal flip, c) random rotation up to 15°C, and d) normalization.

Table 3. Mean and standard deviation of Top-1 test accuracy on CIFAR10 dataset, applying the different optimization methods on 10 randomly drawn training configurations

Architecture	Additive	Multiplicative	Hybrid
	Epoch 5		
ResNet9	61.66 ± 7.47	62.64 ± 9.55	$\mathbf{64.21 \pm 4.85}$
ResNet18	61.16 ± 7.02	60.15 ± 8.30	$\mathbf{63.35 \pm 4.82}$
VGG16	46.49 ± 12.03	$\mathbf{57.72 \pm 18.24}$	56.54 ± 6.31
	Final		
ResNet9	86.24 ± 4.00	85.40 ± 2.18	$\mathbf{86.56 \pm 2.31}$
ResNet18	86.99 ± 3.99	86.23 ± 2.18	$\mathbf{88.33 \pm 2.45}$
VGG16	85.61 ± 6.28	83.00 ± 4.97	$\mathbf{85.67 \pm 4.56}$

In Table 3, the Top-1 average test accuracies and their variance are presented in the 10 different training configurations. More specifically, in rows 3–5, the test accuracies at the fifth epoch are reported, giving us further insight into the training process, while in rows 7–8 the final test accuracies are presented. As demonstrated by the results, the proposed hybrid RMSprop outperforms both the additive and the proposed multiplicative RMSprop. The advantage offered by the multiplicative term is highlighted by the performance of the multiplicative alternative. Even though the multiplicative RMSprop constraints the parameters to their initial sign, it not only achieves a similar test accuracy at the end of the training, but also improves in some cases the test performance during the first epochs. This confirms that multiplicative updates can lead to an acceleration of convergence in the initial stage of training. On the other hand, the proposed hybrid RMSprop sufficiently exploits the advantages of multiplicative update

[1] Xavier initialization randomly draw the weights from a normal distribution defined as $\boldsymbol{w} \sim \mathcal{N}(0, g\sqrt{\frac{2}{M_i N_i}})$, where M_i and N_i are the fan-in and fan-out values of the i-layer, respectively, and g denotes the gain.

term, by combining it with the additive one to overcome the sign limitation, leading to overall better performance in all different architectures.

Table 4. Reports the mean and standard deviation of test accuracy on CIFAR10 dataset applying VGG16, ResNet9 and ResNet18 architectures

Optimizer	VGG16	ResNet9	ResNet18
Fromage [10]	87.31 ± 2.42	87.43 ± 2.19	87.43 ± 2.19
MAdam [11]	73.64 ± 8.17	79.51 ± 10.59	79.51 ± 10.59
Nero [29]	86.45 ± 1.71	86.01 ± 3.78	86.01 ± 3.78
LAMB [30]	82.66 ± 4.65	84.28 ± 5.27	84.28 ± 5.27
SignSGD [31]	78.15 ± 1.78	84.18 ± 1.76	84.18 ± 1.76
Hybrid RMSprop	$\mathbf{88.52 \pm 0.86}$	$\mathbf{87.95 \pm 1.60}$	$\mathbf{90.24 \pm 0.88}$

We extend our experimental setup to evaluate the proposed method also with novel optimization methods which offer robustness and/or minimum to no learning rate tuning over different initial distributions. To evaluate the methods, we randomly draw five configurations, including different gains for the Xavier initialization method and different numbers of training epochs, as already described in the aforementioned experimental setup. For baselines, we used: a) Nero [29] optimizer that applies multiplicative update to perform a projected gradient descent per neuron, using Adam's memory, requiring no learning rate tuning, b) Madam [11] which is also a multiplicative optimizer that allows no learning rate turning and leverages advantages in cases where low-bit width synapses are used, c) Lamb [30] that employs a layerwise adaptive learning rate strategy using normalized gradients, and it is based on Adam optimizer, d) SignSGD [31] that targets distributed training in multiple workers offering compression to gradients and improving the convergence rate of SGD, competing with Adam in particular cases, and e) Fromage [10] that is based on *deep relative trust* controlling the relative size of updates, allowing the training of deep learning models without learning rate tuning.

We report the Top-1 accuracy in the CIFAR10 test set for VGG16, ResNet9, and ResNet18 architectures and in Tables 4. For all the methods evaluated, we use the default hyperparameter configurations. As demonstrated, the proposed hybrid RMSprop outperforms all baselines evaluated in all different cases, regardless of the applied architecture. In this way, we demonstrate that with the proposed hybrid RMSprop, we are able to improve not only the robustness and performance, contrary to the traditional used RMSprop, but also against novel approaches that are applied for robustness. For example, in the ResNet18 case the hybrid RMSprop improves the performance of the models in the average case close to 3% from the second-best approach Fromage.

Furthermore, to investigate the acceleration offered by the proposed approaches, we include the training and test performance in 5 evaluation runs

Table 5. Average and variance of test accuracy over 5 evaluation runs on CIFAR100 using default configurations

Task	Additive	Multiplicative	Hybrid
ResNet18	63.81 ± 1.41	60.18 ± 1.82	$\mathbf{64.68 \pm 1.46}$
ResNet34	64.61 ± 0.51	62.14 ± 0.12	$\mathbf{65.99 \pm 0.44}$

on CIFAR100, which is similar to CIFAR10, except that it has 100 classes containing 600 images each, where 500 are training images and 100 test images per class using the default Xavier initialization ($g = 1$). Similarly to the CIFAR10 task, we applied the default hyperparameters for each optimizer. Cross-entropy loss is applied during training and the Top-1 test accuracy is reported in all cases in Table 5. As also shown in Fig. 2, the multiplicative RMSProp accelerates convergence during the first 10 training epochs. Hybrid RMSprop leads to better accuracy in all cases, effectively exploiting the acceleration capabilities offered by the proposed multiplicative update term. Even though the hybrid

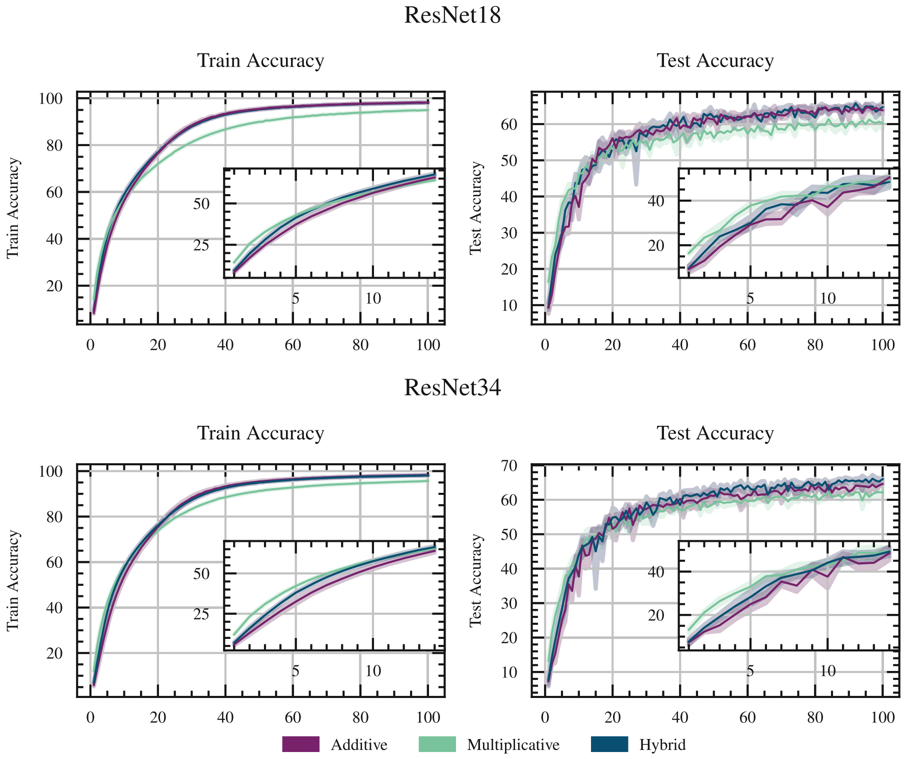


Fig. 2. Training and test accuracy during training applying ResNet18 and ResNet34 architecture on CIFAR100 dataset using default configurations for each optimizer

RMSprop slightly improves the performance in all the aforementioned cases, it consistently leads to improvements on test accuracy, and especially in cases where the robustness is required.

4 Conclusions

Even though multiplicative updates hold theoretical claims that potentially can be used to accelerate DL training and lead to robust models, there are merely studied in the context of DL optimization. In this work, we propose a novel update term that unlocks training acceleration and robust model capabilities, exploiting the properties of multiplicative update rules. More specifically, we propose a multiplicative RMSprop that takes into account the magnitude of parameters, while normalizing gradients, claiming that this leads to faster training and robust models. To overcome the limitations that may arise from the multiplicative update rule, we also propose a hybrid RMSprop that combines the traditional update rule with the proposed multiplicative update term. As demonstrated by the experiments conducted, the proposed methods accelerate training during the initial stage, while ensuring robustness when the actual configurations differ from the initial hypothesis. We first validate the proposed methods on simple convex and non-convex optimization tasks, showing that the benefits observed when the proposed multiplicative term is applied can be generalized in traditionally used image classification benchmarks, such as CIFAR10 and CIFAR100.

Acknowledgments. This project is carried out within the framework of the National Recovery and Resilience Plan Greece 2.0, funded by the European Union - NextGenerationEU (Implementation body: HFRI). Project Number: 16762.

References

1. Brock, A., De, S., Smith, S.L., Simonyan, K.: High-performance large-scale image recognition without normalization (2021). https://arxiv.org/abs/2102.06171
2. Duchi, J., Hazan, E., Singer, Y.: Adaptive subgradient methods for online learning and stochastic optimization. J. Mach. Learn. Res. **12**(7) (2011)
3. Liu, L., et al.: On the variance of the adaptive learning rate and beyond (2019). https://arxiv.org/abs/1908.03265
4. You, Y., Gitman, I., Ginsburg, B.: Large batch training of convolutional networks (2017). https://arxiv.org/abs/1708.03888
5. Sun, S., Cao, Z., Zhu, H., Zhao, J.: A survey of optimization methods from a machine learning perspective (2019)
6. Arora, S., Hazan, E., Kale, S.: The multiplicative weights update method: a meta-algorithm and applications. Theory Comput. **8**(6), 121–164 (2012). https://theoryofcomputing.org/articles/v008a006
7. Littlestone, N.: Learning quickly when irrelevant attributes abound: a new linear-threshold algorithm. Mach. Learn. **2**(4), 285–318 (1988)

8. Lee, D., Seung, H.S.: Algorithms for non-negative matrix factorization. In: Leen, T., Dietterich, T., Tresp, V. (eds.) Advances in Neural Information Processing Systems, vol. 13. MIT Press (2000)
9. Rosenblatt, F.: The perceptron: a probabilistic model for information storage and organization in the brain. Psychol. Rev. **65**(6), 386 (1958)
10. Bernstein, J., Vahdat, A., Yue, Y., Liu, M.Y.: On the distance between two neural networks and the stability of learning. In: Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., Lin, H. (eds.), vol. 33. Curran Associates, Inc., 2020, pp. 21 370–21 381 (2020)
11. Bernstein, J., Zhao, J., Meister, M., Liu, M.Y., Anandkumar, A., Yue, Y.: Learning compositional functions via multiplicative weight updates. In: Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., Lin, H. (eds.) Advances in Neural Information Processing Systems, vol. 33. Curran Associates, Inc., 2020, pp. 13 319–13 330
12. Bachlechner, T., Majumder, B.P., Mao, H., Cottrell, G., McAuley, J.: ReZero is all you need: fast convergence at large depth. In: 37th Conference on Uncertainty in Artificial Intelligence, UAI 2021, no. UAI, pp. 1352–1361 (2021)
13. Zhang, J., He, T., Sra, S., Jadbabaie, A.: Why gradient clipping accelerates training: a theoretical justification for adaptivity (2019). https://arxiv.org/abs/1905.11881
14. Liu, L., et al.: On the variance of the adaptive learning rate and beyond (2019). https://arxiv.org/abs/1908.03265
15. You, Y., Gitman, I., Ginsburg, B.: Large batch training of convolutional networks (2017). https://arxiv.org/abs/1708.03888
16. Kirtas, M., Passalis, N., Mourgias-Alexandris, G., Dabos, G., Pleros, N., Tefas, A.: Robust architecture-agnostic and noise resilient training of photonic deep learning models. IEEE Trans. Emerging Top. Comput. Intell. **7**, 1–10 (2022)
17. Passalis, N., Kirtas, M., Mourgias-Alexandris, G., Dabos, G., Pleros, N., Tefas, A.: Training noise-resilient recurrent photonic networks for financial time series analysis. In: 2020 28th European Signal Processing Conference (EUSIPCO), pp. 1556–1560 (2021)
18. Chorowski, J., Zurada, J.M.: Learning understandable neural networks with nonnegative weight constraints. IEEE Trans. Neural Netw. Learn. Syst. **26**(1), 62–69 (2015)
19. Tsakyridis, A., et al.: Photonic neural networks and optics-informed deep learning fundamentals. APL Photonics **9**(1), 011102 (2024). https://doi.org/10.1063/5.0169810
20. Pappas, C., et al.: A teraflop photonic matrix multiplier using time-space-wavelength multiplexed AWGR-based architectures. In: Optical Fiber Communications Conference and Exhibition (OFC)2024, pp. 1–3 (2024)
21. He, K., Zhang, X., Ren, S., Sun, J.: Delving deep into rectifiers: surpassing human-level performance on ImageNet classification. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV) (2015)
22. Arora, S., Cohen, N., Hazan, E.: On the optimization of deep networks: implicit acceleration by overparameterization. In: Dy, J. Krause, A. (eds.) Proceedings of the 35th International Conference on Machine Learning, ser. Proceedings of Machine Learning Research, vol. 80. PMLR, 10–15 Jul 2018, pp. 244–253 (2018). https://proceedings.mlr.press/v80/arora18a.html
23. Goodfellow, I.J., Vinyals, O., Saxe, A.M.: Qualitatively characterizing neural network optimization problems (2014). https://arxiv.org/abs/1412.6544
24. Paszke, A., et al.: Automatic differentiation in PyTorch (2017)

25. Abadi, M., Agarwal, A., et al.: TensorFlow: large-scale machine learning on heterogeneous systems (2015). https://www.tensorflow.org/, software available from tensorflow.org
26. Hosseini-Asl, E., Zurada, J.M., Nasraoui, O.: Deep learning of part-based representation of data using sparse autoencoders with nonnegativity constraints. IEEE Trans. Neural Netw. Learn. Syst. **27**(12), 2486–2498 (2016)
27. 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)
28. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)
29. Liu, Y., Bernstein, J., Meister, M., Yue, Y.: Learning by turning: neural architecture aware optimisation. In: Meila, M., Zhang, T. (eds.) Proceedings of the 38th International Conference on Machine Learning, ser. Proceedings of Machine Learning Research, vol. 139. PMLR, 18–24 Jul 2021, pp. 6748–6758 (2021)
30. You, Y., et al.: Large batch optimization for deep learning: training BERT in 76 minutes (2020)
31. Bernstein, J., Wang, Y.-X., Azizzadenesheli, K., Anandkumar, A.: signSGD: compressed optimisation for non-convex problems. In: International Conference on Machine Learning, pp. 560–569. PMLR (2018)

TADIL: Task-Agnostic Domain-Incremental Learning Through Task-ID Inference Using Transformer Nearest-Centroid Embeddings

Gusseppe Bravo-Rocca[1](✉), Peini Liu[1], Jordi Guitart[1,2], Ajay Dholakia[3], and David Ellison[3]

[1] Barcelona Supercomputing Center, Barcelona, Spain
{gusseppe.bravo,peini.liu,jordi.guitart}@bsc.es
[2] Universitat Politècnica de Catalunya, Barcelona, Spain
[3] Lenovo Infrastructure Solutions Group, Morrisville, NC, USA
{adholakia,dellison}@lenovo.com

Abstract. Classical Machine Learning models have problems when faced with learning from data that changes over time or across domains due to factors such as noise, occlusion, illumination, or frequency variation, which humans can adapt to without being given independent and identically distributed data. Therefore, a Continual Learning (CL) approach is essential, particularly in this case, the Domain-Incremental Learning. This paper presents a novel pipeline for identifying tasks in domain-incremental learning scenarios without supervision. The pipeline consists of four steps. First, we obtain base embeddings from the raw data through a transformer-based existing model. Second, we group the embedding densities based on similarity and extract few nearest points to each cluster centroid. Third, we train an incremental task classifier using only the previous few points. Finally, we make the pipeline lightweight in terms of computational requirements and build an algorithm that decides when to learn a new task in an online way using the task classifier and drift detector. We experiment with the real-world driving dataset SODA10M and several CL strategies. We conclude that the performance of CL strategies with our pipeline is better not only when the task boundaries are given, but also in the more general practical case of task-agnostic strategies that demand identifying new tasks on-the-fly.

Keywords: Continual learning · Foundation models · Driving dataset · Density clustering

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/978-3-031-78110-0_22.

A. Antonacopoulos et al. (Eds.): ICPR 2024, LNCS 15329, pp. 337–353, 2025.
https://doi.org/10.1007/978-3-031-78110-0_22

1 Introduction

The rapid evolution of Machine Learning (ML) has led to significant advancements in fields like autonomous driving, driven by vast datasets. Traditional ML approaches assume data for training and inference are independent and identically distributed (IID) [5], which rarely holds true in real-world applications. Real-life data can be non-IID, correlated, and context-dependent, leading to the domain shift problem [26]. This is particularly challenging in dynamic environments like autonomous driving, where data distributions drift over time.

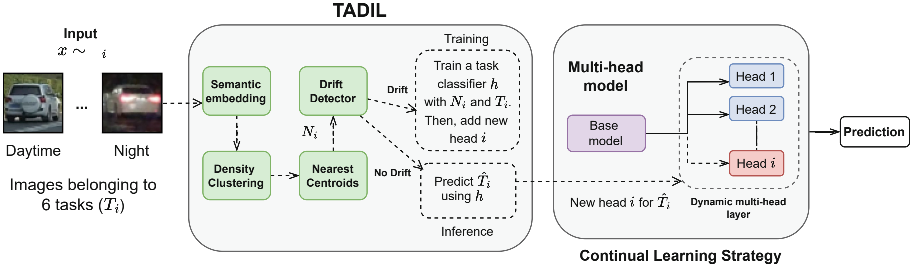


Fig. 1. Workflow of our TADIL method. During training, for a batch x related to task T_i, we derive the nearest-centroid embeddings N_i. If drift is detected, these embeddings are stored in memory $\mathcal{M}$, and a new head is added to the multi-head model, trained on N_i without label supervision. At inference, the task classifier h predicts the task $\hat{T}_i$. The multi-head model then uses the corresponding head for final prediction y. TADIL expands the model with new knowledge during training and accurately selects the relevant head during inference.

Continual Learning (CL) aims to address these challenges by developing algorithms that continuously adapt to new data without forgetting previous knowledge. However, CL faces hurdles such as data privacy, storage, and computational constraints, making it difficult to retain all incoming data for learning. Existing CL strategies, including regularization [28], replay [14], and architectural modifications [19], often assume clear task boundaries and known tasks, which are rare in real-world settings with continuous, unsupervised data. In the context of CL, accurate task identification (task ID) plays a crucial role in maintaining model performance across different domains. Task ID inference allows the model to distinguish between different operational contexts, enabling it to apply the most relevant knowledge and adapt its behavior accordingly. This is particularly important in Domain-Incremental Learning (Domain-IL) scenarios, where the underlying data distribution may shift over time. Without effective task ID inference, a CL system may struggle to differentiate between tasks, leading to suboptimal performance, increased forgetting, and reduced adaptability to new domains. Our work addresses this challenge by introducing a novel, unsupervised approach to task ID inference, which forms the foundation of our Domain-IL pipeline.

In autonomous driving, the complexity increases due to dynamic and unpredictable environments [1]. Vehicles encounter diverse conditions, emphasizing the need for sophisticated Domain-IL approaches. Recognizing these challenges, we introduce a novel, unsupervised Domain-IL pipeline leveraging data embeddings from a pre-trained CLIP model [18] to identify and classify tasks in real-time without explicit task labels. Our approach is tailored for scenarios lacking supervision, utilizing unsupervised clustering of data embeddings to handle dynamic domain shifts.

Our method hypothesizes that semantically similar tasks result in closely clustered embeddings, while dissimilar tasks are farther apart. We extract cluster centroids, representing the most critical data points for training a task-specific classifier, suited for real-world applications like autonomous driving, where accessing all tasks simultaneously is impractical. Our approach adapts to new tasks through semantic embedding, density-based clustering, and a drift detection mechanism, enabling dynamic domain shift detection—a capability yet to be explored within driving datasets.

Implemented as an unsupervised pipeline, our method calculates base embeddings from raw data, groups them by similarity to identify nearest cluster centroids, and incrementally trains a task classifier using these centroids. A lightweight drift detector determines when a new task has emerged, enabling the CL model to adapt to new data domains without explicit supervision. Using the SODA10M real-world driving dataset [7], we demonstrate significant improvements in handling domain shifts and task identification in a task-agnostic manner, with minimal computational overhead.

Contributions:

- Introduce an unsupervised Domain-IL pipeline using nearest-centroid semantic embeddings for task identification and classification in autonomous driving.
- Combine semantic embeddings with density-based clustering and online drift detection for real-time, unsupervised task and domain adaptation.
- Demonstrate the superiority of our method through extensive evaluation on the SODA10M dataset.

2 Related Work

Multimodal Transformer. The integration of multimodal transformers such as CLIP has been pivotal in CL for tasks like autonomous driving. These models effectively combine visual and textual data to generate embeddings that enhance object discrimination and adaptability to new scenarios. For example, the TransFuser model [17] demonstrates improved urban driving predictions by merging image and LiDAR data via attention mechanisms. Additionally, [9] employs a Transformer with multimodal attention to predict vehicular trajectories, acknowledging social interactions. Our methodology leverages pre-trained transformer models to extract environmental patterns without the computational burden of training from scratch, allowing for future updates with more advanced models.

Universal Representation and Cross-Domain Learning. Our work parallels the goals of universal representation learning [12] by creating networks adept at multiple tasks across domains. Unlike the typical use of knowledge distillation in universal representation, which navigates the trade-offs between task-specific losses and gradient conflicts, we utilize the CLIP ViT-B/32 model for direct semantic embedding, streamlining the process. Challenges in task conflict are thus circumvented by our reliance on a singular, robust pre-trained model. In a similar vein, Kim [10] addresses cross-domain adaptability with their Visual Token Matching technique, which, unlike our approach, uses non-parametric patch-level matching for diverse domain learning. Our strategy, while sharing the adaptability objective, diverges in methodology by leveraging the semantic capacity of CLIP ViT-B/32.

Catastrophic Forgetting. Catastrophic forgetting is a major challenge in CL, where neural networks forget previously learned information when acquiring new concepts [2]. This problem, mainly due to gradient descent [11], necessitates balancing the integration of new data (plasticity) and retention of learned knowledge (stability), known as the *stability-plasticity* dilemma. Inspired by biological mechanisms, where the hippocampal system supports rapid learning (plasticity) and the neocortical system maintains long-term storage (stability) [16], various strategies have been proposed. Elastic Weight Consolidation (EWC) [11] protects old knowledge by reducing weight plasticity through regularization, while Experience Replay (ER) [20] mitigates forgetting by maintaining a memory buffer of past experiences combined with new data.

Domain-Incremental Learning (Domain-IL). A subset of CL, focuses on sequentially learning tasks across different domains, essential for recognizing driving patterns under varied conditions. DISC [15] offers an online zero-forgetting solution, learning new domains without re-training and using physical sensors for task ID at inference. Domain-specific autoencoders [4] deduce task IDs through reconstruction error, requiring significant training time and clear task boundaries. In contrast, our approach uses a task-agnostic lightweight classifier learning from small sample sets without predefined boundaries. Domain-aware representations [27] tackle the stability-plasticity issue using a mixture model for incremental learning, adjusting the internal structure to manage drift and imbalance, while our method determines the task ID externally without altering the model's architecture.

Task-Agnostic CL and Task ID Inference. Task-agnostic approaches are crucial for evolving driving scenarios without predefined task identities [29]. Various methods have been proposed to address this challenge. Generative replay [23] retains knowledge of previous tasks but is resource-intensive. Learning without Forgetting (LwF) [13] uses distillation and regularization to preserve knowledge across tasks. Techniques like 'progress and compress' [22] and class-incremental learning [19] avoid task IDs but are tailored to specific models, potentially introducing bias. Recent task ID inference methods, such as Gradient-Based Task Inference [21], use gradient embeddings from model parameters to predict task IDs. However, this approach is computationally intensive and closely tied to specific model architectures. In contrast, our method employs transformer-based semantic embeddings and density-based clustering for task ID inference, offering a more model-agnostic and computationally efficient solution. Our approach uses a lightweight, independent model for task identification, integrating seamlessly with various architectures to ensure unbiased learning and generalization across tasks.

3 Problem Definition

Let X be the input space, Y the output space, and T the space of task IDs. We consider a sequence of K tasks, each with a joint distribution $P_k(X, Y)$ over $X \times Y$. Our goal is to learn a sequence of K models $f_1, f_2, ..., f_K$, where $f_k : X \rightarrow Y$ is the model for task k, such that each model is learned incrementally without forgetting previous tasks.

During inference, the task ID $t \in T$ is unknown. The model f_t predicts the output $y \in Y$ as $f_t = p_t(y|x)$. Formally, our domain-incremental learning approach aims to:

$$\arg \min_{f_1, f_2, \ldots, f_K} \sum_{k=1}^{K} \mathcal{L}(f_k, P_k), \tag{1}$$

where the goal is to minimize the loss function $\mathcal{L}(f_k, P_k)$ for each task k, ensuring incremental learning without forgetting previous knowledge.

In our experiments, the functions $f_1, f_2, \ldots, f_K$ are multi-head classifiers:

$$f_k(\mathbf{x}) = g_k(b(\mathbf{x})), \quad k = 1, 2, \ldots, K, \tag{2}$$

where $b(\mathbf{x})$ is the shared base network, $g_k(\cdot)$ is the classifier for task k, and $\mathbf{x}$ is the input sample. We use the ResNet18 [8] architecture as the base network with linear classifiers.

To improve inference performance and support strategies like EWC, ER, and LwF, a task classifier that learns the task ID without supervision is necessary. This classifier predicts the task ID $t \in T$ from input $\mathbf{x}$, allowing the selection of the appropriate classifier f_t for inference. Without this, the multi-head classifier may suffer from catastrophic forgetting and inefficient knowledge utilization.

Let g_t be the classifier for task t. We define a task classifier $h_t : X \rightarrow T$ that predicts the task ID t for input $\mathbf{x} \in X$. During inference, given $\mathbf{x}$ and the predicted task ID $\hat{t} = h(\mathbf{x})$, the multi-head classifier $f_{\hat{t}}$ predicts the output $y \in Y$:

$$y = f_{\hat{t}}(\mathbf{x}) = g_{\hat{t}}(b(\mathbf{x})). \tag{3}$$

Incorporating a task classifier into the domain-incremental learning framework can be expressed as:

$$\arg \min_{h, g_1, g_2, \ldots, g_K} \sum_{k=1}^{K} \mathcal{L}(f_k, P_k), \tag{4}$$

where h is the task classifier, g_k is the classifier for task k, and $\mathcal{L}(f_k, P_k)$ is the loss function for task k. This ensures each classifier is learned incrementally without forgetting previous knowledge, integrating the task classifier into the learning process.

4 Components of the Pipeline for Task-Agnostic Domain-IL

In this section, we first present a training pipeline for task-agnostic Domain-IL. A series of components are introduced to obtain the nearest centroids and to train an incremental task classifier using the Nearest Centroid Algorithm [24] (a really light classifier). Then, a drift detector is presented to detect when to train incrementally the task classifier. To train the task classifier in an unsupervised way we need a series of components that together can predict the task at inference time. Let h_t be the task classifier for each task T_t. The task classifier is a function that maps an input $x \in \mathcal{X}$ to a task ID $t \in 1, 2, ..., T$, i.e., $h_t : \mathcal{X} \rightarrow 1, 2, ..., T$. The task classifier is obtained through the following pipeline:

Semantic Embedding. Given a batch of inputs $X = x_1, x_2, ..., x_m$, where m is the batch size, we first obtain their corresponding embeddings $E = e_1, e_2, ..., e_m$ using the pretrained transformer-based model CLIP ViT-B/32. We can represent this process as $E = f_{emb}(X)$, where f_{emb} is the embedding function. The use of a pretrained transformer-based model can be justified by the fact that these models have already been trained on large amounts of data, and as a result, have learned representations of common things such as pedestrians, cars, buildings, and other objects that are commonly found in driving scenarios. Moreover, they capture higher-level semantic information about the input that can be useful for various downstream tasks, such as classification, clustering, or retrieval.

Density-Based Clustering. Next, we cluster the embeddings E based on their cosine similarity using the DBSCAN density clustering algorithm [3]. Let the resulting clustering labels be $C = c_1, c_2, ..., c_m$, where c_i is the cluster label assigned to the i-th embedding e_i. Let the clustering function be $f_{clust}(E; \epsilon, minPts)$, where ϵ is the maximum distance between two points for them to be considered in the same cluster (in our setup, $\epsilon = 0.3$) and $minPts$ is the minimum number of points required to form a dense region (in our setup, $minPts = 10$). An example of the outcomes of the density-based clustering can be appreciated in Fig. 2, which depicts the 2D projection of two clusters of embeddings (with actual 512 dimensions) corresponding to two different tasks. The choice of DBSCAN is due to its ability to identify clusters without needing the number of clusters. However, other clustering algorithms can also be used.

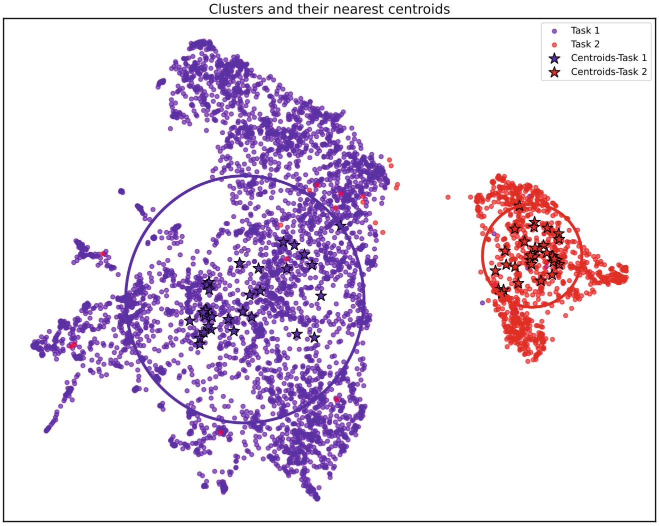


Fig. 2. Visualization of density clustering for task-specific embeddings: Each point represents an embedding, with distinct colors for Task 1 (purple) and Task 2 (red). Black and red stars mark the centroids for each task's cluster. The large circles delineate the areas containing the nearest neighbors to these centroids, emphasizing the core regions of each cluster. This graphical representation highlights the separation between clusters of different tasks and the concentration of embeddings around their central points. (Color figure online)

Nearest-Cluster Centroids. Then, we obtain the nearest centroids of the j distinct clusters present in C ($M = m_1, \ldots, m_j$). Each centroid m_i is calculated in the first phase of the Nearest Centroids Algorithm [24] from all the embeddings e_x where the cluster label $c_x = i$. We can represent this process as $M = f_{cent}(E, C)$, where f_{cent} is the function that obtains the centroids, in our case, by using Manhattan distances. The Manhattan distance metric is employed for centroid computation due to its efficiency in high-dimensional spaces, as commonly encountered in embedding vectors. This metric is faster in those scenarios since we do not have to square the differences as in Euclidean distance. At this point, we obtain the k nearest neighbors of each centroid m_i from the embeddings E using a nearest-neighbor algorithm such as k-Nearest Neighbors (in our setup, $k = 10$). An example of the nearest-cluster centroids can be appreciated in Fig. 2, which depicts the nearest neighbors of each centroid (inside the circles) on top of the clusters corresponding to two different tasks.

Nearest-Centroid Incremental Classifier. Finally, we obtain the task classifier h_t for task T_t by running the second phase of the Nearest Centroids Algorithm. Specifically, given the set of k nearest neighbors of each centroid m_i (let it be N_i), we train a task classifier h_t^i using the nearest neighbors N_i and their corresponding task IDs for tasks $T_1, T_2, ..., T_{t-1}$. The final task classifier h_t for task T_t is obtained by combining the individual classifiers h_t^i using a majority vote. We can represent this process as $h_t = f_{cls}(M_{t^d}, t^d)$, where f_{cls} is the function that obtains the task classifier h_t using the nearest centroids M_{t^d} and t^d being the new task ID detected by the drift detector R (as defined below). Each M_{t^d} is obtained using the training data $D_{i\,i=1}^{t-1}$ of the previously seen tasks. The task classifier h_t maps an input x to a predicted task ID $\hat{t}$, i.e., $h_t(x) = \hat{t}$.

Our classifer excels in tasks with both new and familiar categories, efficiently handling overlapping classes across domains. For instance, in daytime and nighttime scenarios, it accurately categorizes common objects like cars and trucks. TADIL's performance is tuned through key hyperparameters in its clustering (DBSCAN: eps=0.3, min_samples=10) and nearest neighbor (n_neighbors=10) components, balancing overfitting and generalization. This classifier with Manhattan distance proves effective and robust in high-dimensional spaces, though potentially sensitive to feature scaling. To manage multi-head classifier growth in frequent concept drift scenarios, we propose implementing a maximum head limit or merging similar heads based on centroid proximity, maintaining efficiency while adapting to new tasks.

Drift Detector. Additionally, in real scenarios, we need a way to decide when to update incrementally the task classifier h_t as new tasks arrive, that is to say, the trigger t^d, which will allow for effective learning in a domain with a changing task distribution. In order to detect drift between a pair of tasks T_t and $T_{t'}$ over time, we define a drift function R that measures the dissimilarity between the nearest neighbors at different time points:

$$R(N_t, N_{t'}) = \frac{1}{k} \sum_{s=1}^{k} d(N_{t[s]}, N_{t'[s]}) \tag{5}$$

where N_t and $N_{t'}$ are the sets of k nearest neighbors obtained from the embeddings for tasks T_t and $T_{t'}$, respectively. $d(a, b)$ represents the amount of drift between points a and b. This function computes the average drift between the k nearest neighbors in N_t and $N_{t'}$ using the Maximum Mean Discrepancy (MMD) method [6]. A larger value for the drift function indicates a greater difference between the nearest neighbors, suggesting a possible shift in the data distribution, hence, a new task. Because different tasks involve images from different domains, we inevitably get different neighbors, even if the CLIP encoder remains unchanged.

5 Online Algorithm for Task-Agnostic Domain-IL

We present a pipeline algorithm for task-agnostic Domain-IL in an online fashion, utilizing the components introduced earlier (detailed in Algorithm 1).

For each incoming image batch, our algorithm calculates the nearest-centroid embeddings N_t and checks for drift against known tasks stored in memory. Drift is evaluated using the drift detector. If drift is detected, indicating a new task, N_t is saved in memory $\mathcal{M}$, the task classifier h_t is incrementally trained with the new task ID T_t without supervision, and a new head is added to the multi-head classifier for inference until a domain change occurs. If no drift is detected, the classifier h_t estimates the task ID, and if it matches a known task, the model proceeds with inference. If the task ID is unrecognized, it triggers an incremental training phase, updating the task classifier and refining the base model. This adaptive approach relies on the availability of ground truth labels for accurate updates and learning. Additional checks help identify issues with the drift detector or task classifier, ensuring task ID consistency with the multi-head classifier for accurate inference and alignment.

Algorithm 1: Online Task-Agnostic Algorithm for Domain-IL

```
Input: $\mathcal{M}$, memory buffer; $D_t$, current data batch;
Function Online_TADIL($\mathcal{M}, D_t$);
$N_t \leftarrow$ Get_Nearest_Centroid_Embeddings($D_t$);
for $N_{t'}$ in $\mathcal{M}$.Reversed() do
    if not R($N_t$, $N_{t'}$) then
        Use task classifier $h_{t'}$ to predict task ID $T_t$;
        if $T_t \neq T_{t'}$ then
            Raise warning;
        end
        Use head $g_{t'}(D_{t'})$ from the classifier for inference;
        return;
    end
end
Save $N_t$ into $\mathcal{M}$; Train $h_t$ with $N_t$ and $T_t$; Add head $g_t(D_t)$; Use head $g_t(D_t)$;
return;
End Function;
Output: Multi-head classifier updated with new data and task;
```

6 Experimental Evaluation

We used the CLIP model for zero-shot transfer (embeddings of each object category). Besides, our classifier is built upon the ResNet18 model, enhanced for multitasking with a MultiHeadClassifier replacing the original fully connected layer, a linear layer tailored for our specific tasks, and employs an Adam optimizer for training efficiency. The SODA10M dataset was selected for its classification challenges, providing a robust testing ground.

6.1 Testbed

Our experimental setup includes an Ubuntu 22.04 (64-bit) platform, Dual Intel Xeon Platinum 8360Y CPUs @ 2.40 GHz, and 256 GB RAM. The software stack utilizes Docker image intel/oneapi-aikit[1], avalanche-lib 0.3.1[2], torch 1.12.0

[1] https://hub.docker.com/r/intel/oneapi-aikit.

[2] https://avalanche.continualai.org/.

and torchvision 0.13.0[3], intel-extension-for-pytorch 1.12.100+cpu[4], scikit-learn 1.2.2[5], and scikit-learn-intelex 2023.0.1[6]. We focus on the SODA10M dataset [7], adapted for the CLAD-C online classification challenge [25], featuring 20,000 labeled images across various conditions (Fig. 3). Tasks are split into training (80%) and testing (20%) for domain incremental classification, with relevant metrics. For further details on our experimental setup, including the code and scripts used to conduct the experiments, please refer to the supplementary material.

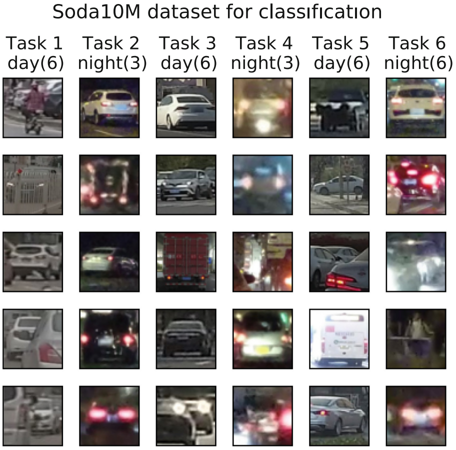


Fig. 3. SODA10M dataset for the CLAD-C benchmark. It consists of six distinct tasks (day/night conditions with different weather, location, cities, and objects), each featuring a specific number of classes (Pedestrian, Cyclist, Car, Truck, Tram/Bus, and Tricycle). For example, Task 1 involves images of six classes captured during the day on city streets in Guangzhou under overcast and rainy weather conditions. Similarly, Task 2 includes images belonging to at most three classes taken at night on highway in Shenzhen under overcast weather condition. The objective of the multi-head model is to classify images accurately for each individual task.

6.2 Performance of the Task Classifier

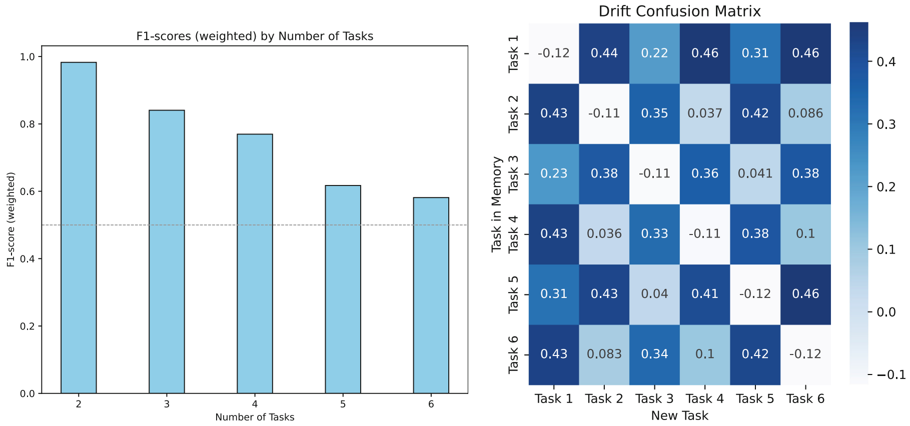
We assess our task classifier's performance by its differentiation among 2 to 6 tasks, as illustrated by Fig. 4a showing F1-scores weighted by task frequency. The F1-score, a harmonic mean of precision and recall, effectively measures performance in contexts with uneven class distributions. As expected, the F1-score

[3] https://www.pytorch.org.
[4] https://github.com/intel/intel-extension-for-pytorch.
[5] https://scikit-learn.org/stable/.
[6] https://github.com/intel/scikit-learn-intelex.

decreases with more tasks, due to the classification problem's increased complexity. However, all tasks score above a 0.5 F1-score, as shown by the horizontal dashed line, indicating consistent, non-random performance. This is further enhanced by using majority voting for batch predictions, reducing the impact of individual errors and not showing a systematic bias towards any class. Thus, the majority voting potentially neutralizes random errors, underscoring the effectiveness of a weighted F1-score approach.

6.3 Performance of the Drift Detector

In this section, we evaluate the performance of our drift detector component. Given a memory component that contains nearest-centroids embeddings from the six tasks, we simulate the arrival at inference time of other embeddings from the same tasks. The objective is to see if the drift detector is able to detect the change of boundaries between tasks. Figure 4b shows the performance of the drift detector by building a confusion matrix to measure the average drift between the k neighbors of each pair of tasks, that is, their dissimilarity. Negative values indicate there is no drift, whereas positive ones indicate drift. As shown in the matrix, drift is correctly detected every time the new task differs from a former task in the memory (and only in this case).



(a) F1-scores for different tasks

(b) Performance of the drift detector

Fig. 4. (a) The bar plot delineates the F1-scores corresponding to number of tasks from 2 to 6. The horizontal dashed line marks the minimum acceptable F1-score of 0.5 (at least not random), with all tasks surpassing this benchmark and (b) depicts the performance of the drift detector, showing the level of drift across a sequence of tasks. Negative values indicate that there is no drift, whereas positive ones indicate drift. As shown, the drift detection is 100% accurate between all the different pairs of tasks.

6.4 Performance of the CL Multi-head Models

This study evaluates the importance of labeling new tasks in multi-head models for accurate task ID identification, crucial for task differentiation and learning.

Task-Boundary Setup. We assess the effectiveness of various CL strategies within a classic CL setup with defined task boundaries. Data is segmented into distinct tasks presented sequentially to the model (no need for drift detection in this case). Our multi-head model, using an Adam optimizer (learning rate 0.01) and cross-entropy loss, tests EWC [11], ER [20], and LwF strategies. Each strategy runs over 4 epochs with a batch size of 200, using defaults from the Avalanche library. Task IDs facilitate model adaptation to different outputs. We compare the effect of both having or not a task ID: *TADIL* (our approach) that predicts the task ID, and a *NoID* approach without task IDs. Table 1 shows that our method surpasses the *NoID* approach in average by efficiently using task IDs (100% task ID match with ground truth). Task IDs enhance the multi-head model's ability to differentiate tasks and apply relevant knowledge, with varying impact across EWC, ER, and LwF. LwF benefits significantly from task IDs due to its use of soft targets, unlike EWC and ER, which balance past and new task learning.

While these strategies aim to mitigate forgetting, task similarity disparities can affect performance. For example, EWC's performance declines when transitioning from day (Task 1, 3 and 5) to night (Task 2, 4 and 6) imagery, which improves with similar subsequent tasks.

Task-Agnostic Setup. The traditional task-boundary setup in CL benchmarks often fails to reflect real-world applications due to the need for explicit task boundaries and the assumption of static, non-overlapping classes. Real-world data streams are dynamic and overlapping. To address this, we evaluate CL strategies (EWC, ER, LwF) in a task-agnostic framework using a multi-head model and detecting new tasks without prior knowledge of task boundaries. The drift detector (Eq. 5) identifies task shifts, triggering training upon new task detection, providing a realistic approach to task identification.

A fair comparison requires periodic retraining of the *NoID* approach in the absence of task IDs, particularly at the start of daytime tasks (Tasks 1, 3, and 5), simulating practical retraining scenarios to adapt to new data distributions. As shown in Table 1, "Without repetitions" column shows, in average, improved accuracy across all strategies with task ID provision, validating our task classifier and drift detector's efficacy. In some cases, for Task 1, retraining works better that having a task ID, however, it can not generalize for the rest of tasks, as we can see for EWC and LwF strategies. The *TADIL* approach performs consistently across task-boundary and task-agnostic scenarios due to accurate drift detection, while the *NoID* approach shows performance disparity, with scheduled retraining mitigating task forgetting for daytime tasks.

This highlights the trade-off between model update frequency and task forgetting, with *TADIL* maintaining superiority over the *NoID* approach. Notably, the Replay strategy exhibits the lowest task forgetting, emphasizing the effectiveness of our method in managing model updates.

Task-Agnostic Setup with Task Repetitions. We evaluate CL strategies (EWC, ER, LwF) in scenarios with repeating tasks, mimicking real-world conditions with varying environments. A custom sequence [1, 2, 3, 2, 4, 4, 5, 5, 5, 6] tests the model's adaptability. Our *TADIL* method outperforms the *NoID* method, especially for nighttime tasks (Table 1). Unlike previous experiments, the *NoID* approach's performance is significantly affected by irregular retraining intervals, either due to prolonged periods without updates or consecutive updates for repeated tasks. Continual retraining on repeated tasks (e.g., Task 5) without task detection can lead to redundancy, inefficiency, and potential overfitting, as seen with the diminishing returns of the LwF strategy after successive retrainings on Task 5. This issue is critical in resource-constrained settings, emphasizing

Table 1. Average accuracy and standard errors for each task, CL strategy, and scenario. Bold indicates best accuracy. The NoID method sometimes generalizes better for Task 1, but struggles for other tasks.

		With boundaries		Without repetitions		With repetitions	
Strategy	Task	TADIL	NoID	TADIL	NoID	TADIL	NoID
EWC	1	**0.63** ±.04	0.63 ±.04	0.60 ±.05	**0.70** ±.05	0.60 ±.05	**0.69** ±.04
	2	**0.82** ±.05	0.77 ±.11	**0.83** ±.05	0.75 ±.07	**0.85** ±.04	0.75 ±.07
	3	**0.71** ±.03	0.59 ±.08	0.66 ±.06	**0.68** ±.03	0.66 ±.08	**0.69** ±.04
	4	**0.88** ±.02	0.79 ±.09	**0.85** ±.04	0.75 ±.07	**0.87** ±.03	0.75 ±.06
	5	**0.79** ±.02	0.72 ±.05	**0.79** ±.02	0.76 ±.01	**0.77** ±.03	0.77 ±.02
	6	**0.78** ±.03	0.67 ±.05	**0.79** ±.01	0.57 ±.08	**0.73** ±.04	0.59 ±.04
	Avg	**0.77** ±.04	0.70 ±.07	**0.75** ±.04	0.70 ±.05	**0.75** ±.04	0.71 ±.05
LwF	1	**0.56** ±.04	0.56 ±.04	0.57 ±.03	**0.66** ±.06	0.57 ±.05	**0.64** ±.06
	2	**0.88** ±.02	0.81 ±.06	**0.85** ±.04	0.75 ±.05	**0.89** ±.01	0.81 ±.06
	3	**0.70** ±.03	0.49 ±.03	0.67 ±.02	**0.68** ±.06	**0.69** ±.04	0.64 ±.06
	4	**0.88** ±.01	0.78 ±.05	**0.86** ±.03	0.74 ±.06	**0.88** ±.01	0.79 ±.07
	5	**0.75** ±.01	0.49 ±.07	**0.73** ±.02	0.76 ±.06	**0.75** ±.03	0.72 ±.07
	6	**0.69** ±.01	0.56 ±.05	**0.68** ±.02	0.51 ±.03	**0.70** ±.01	0.48 ±.07
	Avg	**0.74** ±.03	0.61 ±.05	**0.73** ±.03	0.68 ±.05	**0.75** ±.03	0.68 ±.06
Replay	1	**0.68** ±.04	0.68 ±.04	**0.67** ±.04	0.64 ±.06	0.69 ±.03	**0.70** ±.04
	2	**0.89** ±.02	0.82 ±.05	**0.88** ±.02	0.75 ±.08	**0.89** ±.02	0.78 ±.05
	3	**0.75** ±.01	0.70 ±.03	**0.76** ±.02	0.61 ±.06	**0.77** ±.02	0.70 ±.03
	4	**0.89** ±.01	0.79 ±.03	**0.89** ±.01	0.77 ±.07	**0.88** ±.01	0.79 ±.03
	5	**0.82** ±.01	0.73 ±.02	**0.83** ±.01	0.76 ±.06	**0.83** ±.01	0.80 ±.01
	6	**0.83** ±.01	0.76 ±.05	**0.83** ±.02	0.58 ±.06	**0.83** ±.01	0.59 ±.06
	Avg	**0.81** ±.02	0.75 ±.04	**0.81** ±.02	0.68 ±.06	**0.81** ±.02	0.72 ±.04

the need for task detection mechanisms to optimize continual learning. Our findings demonstrate that our approach maintains superior performance in complex scenarios with task repetitions, highlighting its robustness and applicability in real-world challenges.

6.5 Forgetting Rate Analysis

To quantify TADIL's effectiveness in mitigating catastrophic forgetting, we calculated the overall average forgetting rate for both TADIL and the NoID baseline across all scenarios and tasks using:

$$\bar{F} = \frac{1}{N_{\text{scenarios}}} \sum_{e=1}^{N_{\text{scenarios}}} \frac{1}{N_{\text{tasks}} - 1} \sum_{t=1}^{N_{\text{tasks}}-1} \Delta\text{Acc}_{t,t+1}^{\text{method, e}} \tag{6}$$

where $\Delta\text{Acc}_{t,t+1}^{\text{method, e}}$ represents the change in accuracy between consecutive tasks t and $t+1$ for the method in scenario e, $N_{\text{scenarios}}$ is the number of scenarios, and N_{tasks} is the total number of tasks. Table 2 shows the overall average forgetting rates for TADIL and NoID methods. TADIL demonstrates a negative forgetting rate (–0.0054), indicating improved performance on previous tasks as it learns new ones. In contrast, NoID exhibits a positive forgetting rate (0.0275), suggesting a decline in performance on previously learned tasks.

Table 2. Forgetting rate across the 6 tasks for TADIL and NoID methods.

Method	Task 1–2	Task 2–3	Task 3–4	Task 4–5	Task 5–6	Average
TADIL	–0.14	0.05	–0.08	0.08	0.03	–0.0054
NoID	–0.03	0.12	–0.07	0.05	0.06	0.0275

The negative forgetting rate suggests that TADIL leverages knowledge from new tasks to refine its understanding of previous tasks, a highly desirable characteristic in CL scenarios.

7 Conclusion

In this paper, we proposed a novel pipeline called TADIL for detecting and identifying tasks in task-agnostic Domain-IL scenarios without supervision. Our pipeline first obtains base embeddings from the raw data using an already existing transformer-based model. The embedding densities are grouped based on their similarity to obtain the nearest points to each cluster centroid and a task classifier is incrementally trained using only these few points. This task classifier and a drift detector are used together to learn new tasks. Our experiments using the SODA10M real-world driving dataset have demonstrated the good performance of the drift detector and the task classifier, and how state-of-the-art CL strategies work when using our pipeline to predict the task ID, both in experiments assuming task boundaries using a traditional approach, and also in more realistic task-agnostic scenarios that require detecting new tasks on-the-fly.

Limitations and Future Work. The current limitation of the proposed approach is its dependence on pre-trained models for embedding extraction, which may not perform well with rare or domain-specific elements, such as emergency vehicles without visible signals or non-standard traffic signs in driving datasets. Future work will focus on model fine-tuning using representative datasets of these uncommon classes to improve task ID identification and robustness in real-world conditions. Additionally, plans include developing a custom ER strategy using zero-shot predictions to generate weak labels for centroids, facilitating unsupervised training and optimizing the CL strategy.

Supplementary Material. For further details on our experimental setup, including the code and scripts used to conduct the experiments, please refer to our Github repository[7] and supplementary file.

Acknowledgments. We thank Lenovo for providing the technical infrastructure to run the experiments in this paper. This work was partially supported by Lenovo and Intel as part of the Lenovo AI Innovators University Research program, by the Spanish Government under contract PID2019-107255GB-C22, and by the Generalitat de Catalunya under contract 2021-SGR-00478

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article

References

1. Chen, L., Wu, P., Chitta, K., Jaeger, B., Geiger, A., Li, H.: End-to-end autonomous driving: challenges and frontiers (2023)
2. De Lange, M., et al.: A continual learning survey: defying forgetting in classification tasks. IEEE Trans. Pattern Anal. Mach. Intell. **44**(7), 3366–3385 (2022). https://doi.org/10.1109/TPAMI.2021.3057446
3. Ester, M., Kriegel, H.P., Sander, J., Xu, X., et al.: A density-based algorithm for discovering clusters in large spatial databases with noise. In: Kdd, vol. 96, pp. 226–231 (1996)
4. González, C., Sakas, G., Mukhopadhyay, A.: What is wrong with continual learning in medical image segmentation? (2020). https://arxiv.org/abs/2010.11008
5. Goodfellow, I., Bengio, Y., Courville, A.: Deep learning. MIT Press (2016). https://www.deeplearningbook.org/
6. Gretton, A., Borgwardt, K.M., Rasch, M.J., Schölkopf, B., Smola, A.: A kernel two-sample test. J. Mach. Learn. Res. **13**(25), 723–773 (2012). http://jmlr.org/papers/v13/gretton12a.html
7. Han, J., et al.: SODA10M: A Large-Scale 2D Self/Semi-Supervised Object Detection Dataset for Autonomous Driving (2021), arXiv preprint arXiv:2106.11118

[7] https://github.com/gusseppe/TADIL.

8. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proc. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2016), 27–30 June, pp. 770–778 (2016). https://doi.org/10.1109/CVPR.2016.90
9. Huang, Z., Mo, X., Lv, C.: Multi-modal motion prediction with transformer-based neural network for autonomous driving. In: Proceedings of 39th International Conference on Robotics and Automation (ICRA), May 23–27, pp. 2605–2611 (2022). https://doi.org/10.1109/ICRA46639.2022.9812060
10. Kim, D., Kim, J., Cho, S., Luo, C., Hong, S.: Universal few-shot learning of dense prediction tasks with visual token matching (2023), arXiv preprint arXiv:2303.14969
11. Kirkpatrick, J., et al.: Overcoming catastrophic forgetting in neural networks. Proc. Natl. Acad. Sci. **114**(13), 3521–3526 (2017). https://doi.org/10.1073/pnas.1611835114
12. Li, W.H., Liu, X., Bilen, H.: Universal representations: A unified look at multiple task and domain learning (2022), arXiv preprint arXiv:2204.02744
13. Li, Z., Hoiem, D.: Learning without forgetting (2017). https://arxiv.org/abs/1606.09282
14. Lopez-Paz, D., Ranzato, M.: Gradient episodic memory for continual learning. In: Advances in Neural Information Processing Systems, vol. 30 (NIPS 2017), pp. 6470–6479. Curran Associates Inc. (2017)
15. Mirza, M., Masana, M., Possegger, H., Bischof, H.: An efficient domain-incremental learning approach to drive in all weather conditions. In: Proc. 2022 IEEE/CVF Conf. on Computer Vision and Pattern Recognition Workshops (CVPRW 2022), Jun 19–20, pp. 3000–3010 (2022). https://doi.org/10.1109/CVPRW56347.2022.00339
16. Parisi, G.I., Kemker, R., Part, J.L., Kanan, C., Wermter, S.: Continual lifelong learning with neural networks: a review. Neural Netw. **113**, 54–71 (2019). https://doi.org/10.1016/j.neunet.2019.01.012
17. Prakash, A., Chitta, K., Geiger, A.: Multi-modal fusion transformer for end-to-end autonomous driving. In: Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR 2021), Jun 19–25, pp. 7073–7083. IEEE Computer Society (2021). https://doi.org/10.1109/CVPR46437.2021.00700
18. Radford, A., et al.: Learning transferable visual models from natural language supervision. In: Proceedings of 38th International Conference on Machine Learning (ICML 2021). Proceedings of Machine Learning Research, Jul 18–24, vol. 139, pp. 8748–8763. PMLR (2021)
19. Rebuffi, S.A., Kolesnikov, A., Sperl, G., Lampert, C.H.: iCaRL: incremental classifier and representation learning. In: Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2017), 21–26 Jul, pp. 5533–5542 (2017). https://doi.org/10.1109/CVPR.2017.587
20. Rolnick, D., Ahuja, A., Schwarz, J., Lillicrap, T., Wayne, G.: Experience Replay for Continual Learning. In: Advances in Neural Information Processing Systems (NeurIPS 2019), vol. 32. Curran Associates, Inc. (2019)
21. Roy, S., Verma, V.K., Gupta, D.: Efficient expansion and gradient based task inference for replay free incremental learning (2023). https://arxiv.org/abs/2312.01188
22. Schwarz, J., et al.: Progress & compress: a scalable framework for continual learning. In: Proceedings 35th International Conference on Machine Learning (ICML 2018). Proceedings of Machine Learning Research, 10–15 July vol. 80, pp. 4535–4544. PMLR (2018)

23. Shin, H., Lee, J.K., Kim, J., Kim, J.: Continual learning with deep generative replay. In: Advances in Neural Information Processing Systems, vol. 30 (NIPS 2017), pp. 2994–3003. Curran Associates Inc. (2017)
24. Tibshirani, R., Hastie, T., Narasimhan, B., Chu, G.: Diagnosis of multiple cancer types by shrunken centroids of gene expression. Proc. Natl. Acad. Sci. U.S.A. **99**(10), 6567–6572 (2002)
25. Verwimp, E., et al.: CLAD: a realistic continual learning benchmark for autonomous driving. Neural Netw. **161**, 659–669 (2023). https://doi.org/10.1016/j.neunet.2023.02.001
26. Wang, M., Deng, W.: Deep visual domain adaptation: a survey. Neurocomputing **312**, 135–153 (2018). https://doi.org/10.1016/j.neucom.2018.05.083, https://arxiv.org/abs/1802.03601
27. Xie, J., Yan, S., He, X.: General incremental learning with domain-aware categorical representations. In: Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR 2022), 21–24 Jun, pp. 14331–14340. IEEE Computer Society (2022). https://doi.org/10.1109/CVPR52688.2022.01395
28. Zenke, F., Poole, B., Ganguli, S.: Continual learning through synaptic intelligence. In: Proceedings of 34th International Conference on Machine Learning (ICML 2017), 6–11 Aug. Proceedings of Machine Learning Research, vol. 70, pp. 3987–3995. PMLR (2017)
29. Zhu, H., Majzoubi, M., Jain, A., Choromanska, A.: TAME: Task Agnostic Continual Learning using Multiple Experts, arXiv preprint arXiv:2210.03869 (2022)

Evidential Federated Learning for Skin Lesion Image Classification

Rutger Hendrix[1(✉)], Federica Proietto Salanitri[1], Concetto Spampinato[1], Simone Palazzo[1], and Ulas Bagci[2]

[1] Department of Electrical, Electronic and Computer Engineering,
University of Catania, Catania, Italy
rutger.hendrix@phd.unict.it,
{federica.proiettosalanitri,concetto.spampinato,
simone.palazzo}@unict.it

[2] Machine and Hybrid Intelligence Lab, Department of Radiology,
Northwestern University, Chicago, USA
ulas.bagci@northwestern.edu

Abstract. We introduce *FedEvPrompt*, a federated learning approach that integrates principles of evidential deep learning, prompt tuning, and knowledge distillation for distributed skin lesion classification. FedEvPrompt leverages two sets of prompts: *b-prompts* (for low-level basic visual knowledge) and *t-prompts* (for task-specific knowledge) prepended to frozen pre-trained Vision Transformer (ViT) models trained in an evidential learning framework to maximize class evidences. Crucially, knowledge sharing across federation clients is achieved only through knowledge distillation on attention maps generated by the local ViT models, ensuring enhanced privacy preservation compared to traditional parameter or synthetic image sharing methodologies. FedEvPrompt is optimized within a round-based learning paradigm, where each round involves training local models followed by attention maps sharing with all federation clients. Experimental validation conducted in a real distributed setting, on the ISIC2019 dataset, demonstrates the superior performance of FedEvPrompt against baseline federated learning algorithms and knowledge distillation methods, without sharing model parameters. In conclusion, FedEvPrompt offers a promising approach for federated learning, effectively addressing challenges such as data heterogeneity, imbalance, privacy preservation, and knowledge sharing.

Keywords: Prompt Tuning · Knowledge Distillation · Uncertainty

1 Introduction

In recent decades, deep learning has played a leading role in medical image analysis, including skin lesion classification. However, most of the existing methods rely on centralized learning, assuming data uniformity and accessibility, which

A. Antonacopoulos et al. (Eds.): ICPR 2024, LNCS 15329, pp. 354–365, 2025.
https://doi.org/10.1007/978-3-031-78110-0_23

often does not align with the reality of decentralized and privacy-sensitive clinical settings. This disparity not only limits progress in the field, but also exacerbates inequalities, with wealthier regions having a data advantage over poorer areas, leading to disparities in model performance and clinical support. Federated learning (FL) emerges as a promising solution to this challenge, enabling model training across distributed devices while preserving data privacy. Methods like FedAvg [10] and FedProx [8] have addressed issues such as non-i.i.d. data and system heterogeneity, yet they still face obstacles, particularly in scenarios with class imbalances and data heterogeneity. Evidential Deep Learning (EDL) [13] has found adoption in FL to handle these limitations in medical data, thereby enhancing model confidence and reliability, crucial for clinical applications. For example, the recent work on uncertainty-aware aggregation of federated models for diabetic retinopathy classification demonstrates its efficacy in improving model performance and reliability [19].

Furthermore, the scarcity of data poses an additional significant challenge, often leading to model overfitting and suboptimal federation performance. Recent techniques like learnable prompting [9], particularly effective in low-data regimes, offer a promising solution by facilitating personalized model tuning across distributed clients [7]. Nonetheless, privacy concerns persist, particularly due to the sharing and aggregation of model parameters, which poses the risk of reconstructing training images, as demonstrated by recent studies [4,20]. To mitigate these concerns, one strategy involves sharing suitably-constructed synthetic data generated through generative models [11]. Yet, the use of generative models carries its own risks, potentially incorporating and synthesizing sensitive training samples, thus exacerbating privacy concerns.

We here propose FedEvPrompt, a novel approach that integrates principles of evidential deep learning, prompt tuning, and knowledge distillation to address existing limitations comprehensively. FedEvPrompt leverages prompts prepended to pre-trained ViT models trained in an evidential learning setting, maximizing class evidence. Knowledge sharing across federation clients is achieved only through knowledge distillation on attention maps generated by ViT models, which offers greater privacy preservation compared to sharing parameters or synthetic images, as it lacks pixel-level details and reconstructive qualities. While our approach maintains a high level of abstraction for minimizing privacy leaks, it also provides richer information than average logits, as in FedDistill [15], or prototypes, as in FedProto [17]. Thus, FedEvPrompt represents a principled way to share insights into the decision-making process of local models for enhanced federated performance, as demonstrated by the results achieved on a real-word distributed setting for skin lesion classification.

2 Background Evidential Learning

Deep Learning methods often use softmax activation in the output layers to perform classification. However, softmax outputs can be biased to training data, failing to predict with low certainty even for samples far from the distribution

[16]. In contrast to the additivity principle in probability theory, Dempster-Shafer theory describes that the sum of belief can be less than 1. Its remainder is then attributed to uncertainty.

In a frame of K mutually exclusive singletons (e.g., class labels), each singleton $k \in [K]$ is assigned a belief mass b_k, and an overall uncertainty mass u. The sum of these $K+1$ mass values is constrained by $u + \sum_{k=1}^{K} b_k = 1$, with $u \geq 0$, $b_k \geq 0, \forall k \in [K]$. The belief mass is determined by the evidence supporting each singleton, reflecting the level of support gathered from data. The uncertainty is inversely proportional to the total amount of evidence, with uncertainty equal to 1 for a total lack of evidence. A belief mass assignment corresponds to a Dirichlet distribution with parameters $\alpha_k = e_k + 1$, where e_k denotes the derived evidence for the k-th singleton. This choice of Dirichlet distribution is motivated by its role as a conjugate prior to the categorical distribution, and is defined as:

$$\mathrm{Dir}(p, \alpha) = \frac{\mathit{\Gamma}(S)}{\prod_{k=1}^{K} \mathit{\Gamma}(\alpha_k)} \prod_{k=1}^{K} p_k^{\alpha_k - 1}, \quad \alpha_k > \ 0$$

where p denotes a probability mass function, K denotes the number of classes, $\alpha = [\alpha_1, \ldots, \alpha_K]$ are the Dirichlet parameters related to the evidence, $\mathit{\Gamma}(\cdot)$ denotes the gamma function, and $S = \sum_{k=1}^{K} \alpha_k$ is termed the Dirichlet strength.

From the parameters of this Dirichlet distribution, the belief b_k and the uncertainty u are derived as:

$$b_k = \frac{\alpha_k - 1}{S}, \quad u = \frac{K}{S}$$

When considering an opinion, the expected probability $\hat{p}_k$ of the k-th singleton equates to the mean of the corresponding Dirichlet distribution, calculated by:

$$\hat{p}_k = \frac{\alpha_k}{S}$$

Although this modeling of second-order probabilities and uncertainty enables the computation of different types of uncertainties, this work only considers classical vacuity uncertainty (u).

Evidential Deep Learning (EDL) aims to quantify these uncertainties in the predictions, using a single deterministic neural network. The model learns evidence from the logit layer, typically applying non-negative functions like ReLU to obtain these values. With these minimal changes, EDL models can be trained by minimizing losses such as evidential mean squared error (MSE) loss to form the multinomial opinions for K-class classification of a given sample i as a Dirichlet distribution. Following [14], the evidential MSE loss for sample i can be interpreted as:

$$\mathcal{L}_i(\mathit{\Theta}) = \int \left[(\mathbf{y}_i - \mathbf{p}_i)^T (\mathbf{y}_i - \mathbf{p}_i) \right] \mathrm{Dir}(\mathbf{p}_i, \alpha_i) d\mathbf{p}_i$$

$$= \sum_{k=1}^{K} \mathbb{E}\left[y_{i,k}^2 - 2y_{i,k}y_{i,k} + p_{i,k}^2\right]$$

where $\mathbf{y}_i = (y_{i,1}, \ldots, y_{i,K})$ and $\mathbf{p}_i = (p_{i,1}, \ldots, p_{i,K})$ are vectors of true and predicted probabilities, and $\alpha_i = (\alpha_{i,1}, \ldots, \alpha_{i,K})$ are the Dirichlet parameters.

Using the identity $\mathbb{E}\left[p_{i,k}^2\right] = \mathbb{E}\left[p_{i,k}\right]^2 + \mathrm{Var}(p_{i,k})$, the loss can be rewritten as:

$$\mathcal{L}_i(\Theta) = \sum_{k=1}^{K} \left((y_{i,k} - \mathbb{E}\left[p_{i,k}\right])^2 + \mathrm{Var}(p_{i,k})\right)$$

$$= \sum_{k=1}^{K} \left(\left(y_{i,k} - \frac{\alpha_{i,k}}{S_k}\right)^2 + \frac{\alpha_{i,k}\left(S_i - \alpha_{i,k}\right)}{S_i^2(S_i + 1)}\right)$$

where S_i is the total Dirichlet strength for sample i.

To avoid generating misleading evidence for incorrect labels, a Kullback-Leibler (KL) divergence regularization term is used, reducing total evidence to zero for incorrectly classified samples. The KL term is defined as:

$$\mathcal{L}_{KL} = \mathrm{KL}\left[\mathrm{Dir}(p_i \mid \widetilde{\boldsymbol{\alpha}}_i) \| \mathrm{Dir}(p_i \mid \mathbf{1})\right]$$

$$= \log\left(\frac{\Gamma\left(\sum_{k=1}^{K} \widetilde{\alpha}_{i,k}\right)}{\Gamma(K)\prod_{k=1}^{K} \Gamma(\widetilde{\alpha}_{i,k})}\right) + \sum_{k=1}^{K} (\widetilde{\alpha}_{i,k} - 1)\left(\psi(\widetilde{\alpha}_{i,k}) - \psi\left(\sum_{k=1}^{K} \widetilde{\alpha}_{i,j}\right)\right)$$

where $\widetilde{\boldsymbol{\alpha}}_i$ are the Dirichlet parameters after the removal of non-misleading evidence defined as $\widetilde{\boldsymbol{\alpha}} = y + (1 - y) \odot \alpha$, KL[.‖.] denotes the Kullback-Leibler divergence operator, and $\psi(.)$ is the digamma function [14]. The final **evidential loss** $\mathcal{L}_\epsilon$ results in:

$$\mathcal{L}_\epsilon = \mathcal{L}_i + \mathcal{L}_{KL}$$

3 Methodology

We introduce FedEvPrompt, our federated learning paradigm, which leverages prompt evidential learning and knowledge distillation on ViT attention maps for enabling effective knowledge aggregation across federated clients. The overall learning strategy is described in Fig. 1.

FedEvPrompt is based on a pre-trained ViT model, kept frozen across all clients within the federation. Upon the fixed backbone, prompts are prepended on each client model and optimized using local data. Each client also computes attention maps (through attention rollout mechanism [1]) for each class and shares a subset of them with the federation. The attention maps by all clients form our *uncertainty-aware attention buffer* that is used for knowledge distillation during prompt learning.

Learning is organized in rounds: at each round, federation clients carry out different local training epochs for prompt optimization through a combination of evidential loss for learning class evidence and knowledge distillation loss on the

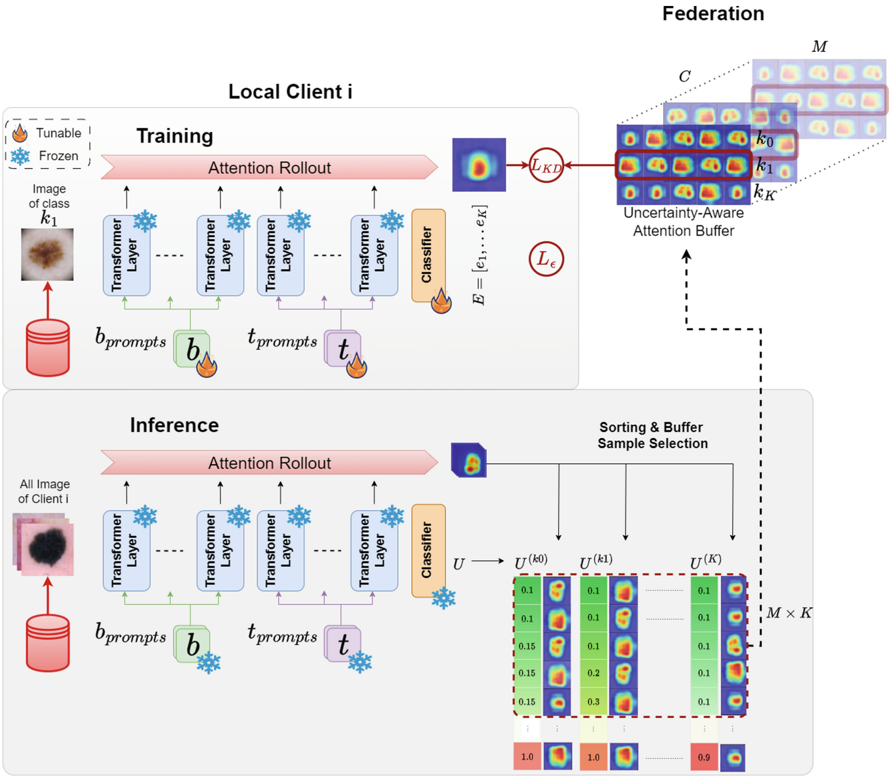


Fig. 1. Overview of FedEvPrompt. During a round of **Training (top)**, local data is used to optimize *b-prompts*, encoding general visual features, and *t-prompts*, encoding task-specific information, prepended to a frozen ViT encoder. Optimization is carried out by minimizing evidential loss ($\mathcal{L}_{\epsilon}$) and a knowledge distillation loss ($\mathcal{L}_{\mathcal{KD}}$) between local attention maps and those of the federation available in the *uncertainty-aware attention buffer*. After a round of training at **Inference (bottom)**, the client identifies, for each class K, its M most informative attention rollout maps (sorted by lowest uncertainty) to contribute to the federated *uncertainty-aware attention buffer*.

per-class attention maps present within the buffer. At the end of training round, each client identifies its M most informative attention maps for each class and updates the buffer.

More in detail, each client employs a frozen ViT, as the backbone, with two sets of prompts *b-prompts* and *t-prompts*. Each prompt is associated with a specific attention layer, with the *b-prompts* (basic prompts) prepended to layers with low-level feature representation and the *t-prompts* (task-specifc prompts) to deeper layers with high-level feature representation. The parameters of the prompts are incorporated through pre-fix tuning. Let's denote the output of the i_{th} attention layer as h_i with $i \in 1 \cdots H$ where H is the number of attention layers. The prepended parameters for the key and value inputs, denoted as pr_k and pr_v respectively, are introduced as follows:

$$MSA(h_Q, [pr_k^{(i)}; h_K], [pr_v^{(i)}; h_V]) \tag{1}$$

where h_Q, h_K, and h_V represent the query, key, and value outputs from the previous layer, respectively. The prepended prompts $pr_k^{(i)}$ and $pr_v^{(i)}$ are *b-prompts* to the first l layers, and *t-prompts* to the last $H - l$ layers. The two sets of prompts undergo distinct optimization strategies: *b-prompts* require slower adaptation since the frozen backbone (ViT) has already grasped general visual features. Conversely, *t-prompts* necessitate faster adjustments to accommodate varying data distributions. Consequently, $\mu_1 < \mu_2$, where μ_1 and μ_2 denote the learning rates for the *b-prompts* and *g-prompts* optimizers, respectively. Both set of prompts are optimized by minimizing an overall loss $\mathcal{L}_G$ that includes an evidential loss term $\mathcal{L}_\epsilon$ and a knowledge distillation loss term $\mathcal{L}_{\mathcal{KD}}$ on the shared attention map buffer A:

$$\mathcal{L}_G = \mathcal{L}_\epsilon + \lambda \mathcal{L}_{\mathcal{KD}} \tag{2}$$

where $\lambda = 1e^{-6}$ is a parameter controlling the balance between the two terms.

3.1 Evidential Loss

Our method is based on evidential learning, i.e., the classification model outputs evidences $E = [e_1, \ldots, e_K]$, with K categorical class elements (number of classes). The Dirichlet distribution characterizes the likelihood of each discrete probability value within a set of possible probabilities. It is parameterized by a vector of K elements (classes), $\boldsymbol{\alpha} = [\alpha_1, \ldots, \alpha_K]$, defined as $\alpha_k = e_k + W_k$, with e_k being the model evidence for class k, and W_k the prior weight for that class. Classical EDL assumes a uniform Dirichlet (Dir(1)) distribution as a prior, i.e., $\boldsymbol{W} = \langle 1, 1, \ldots, 1 \rangle$. The uncertainty for the i^{th} input sample is then estimated as $u_i = \frac{K}{S}$, with S being the total Dirichlet strength $S = \sum_{k=1}^{K} \alpha_k$. Due to the strong class imbalance typical in federated learning settings, we change the uniform evidential prior to a skewed distribution, weighted by class frequency:

$$\alpha_k = e_k + W_k \quad \text{with} \quad W_k = \frac{K}{K-1}\left(1 - \frac{N_k}{N}\right) \tag{3}$$

such that $\sum_{k=1}^{K} W_k = K$.

Prompt parameters are finally optimized by minimizing the evidential loss defined in [14] as a combination of MSE and KL divergence. Given the i^{th} input sample, the one-hot-encoded vector y_i of its class label k, and its expected probabilities p_i, the **evidential loss** $\mathcal{L}_\epsilon$ is computed as:

$$\mathcal{L}_\epsilon(\boldsymbol{\theta}) = \mathbb{E}_{\mathbf{p} \sim \mathrm{Dir}(\boldsymbol{\alpha})}\left[(y_i - p_i)^T (y_i - p_i)\right] + \lambda_{KL} D_{KL}(\mathrm{Dir}(p_i | \widetilde{\alpha}_i) || \mathrm{Dir}(p_i | \mathbf{w}_i)) \tag{4}$$

with $\lambda_{KL} = min(1, t/10)$ being an annealing factor applied to gradually increase the regularization impact with the number of epochs t.

In order to let the evidence for incorrect classes shrink to the weighted prior values $\mathbf{W}$, the KL divergence loss term minimizes the distributional difference between $\mathbf{W}$ and misleadingevidence $\widetilde{\boldsymbol{\alpha}}$, formulated as $\widetilde{\boldsymbol{\alpha}} = y \cdot \mathbf{w} + (1 - y) \odot \boldsymbol{\alpha}$.

Given the weighted prior distribution $\mathbf{W} = \mathrm{Dir}(\mathbf{p}|\mathbf{w})$ and $P = \mathrm{Dir}(\mathbf{p}|\widetilde{\boldsymbol{\alpha}})$, the general KL divergence form for the Dirichlet distribution [12] becomes:

$$\begin{aligned} D_{KL}(\mathrm{Dir}(p|\widetilde{\alpha})||\mathrm{Dir}(p|\mathbf{w})) = \log\left(\frac{\Gamma\left(\sum_{k=1}^{K}\widetilde{\alpha}_k\right)}{\Gamma\left(\sum_{k=1}^{K}w_k\right)}\right) + \sum_{k=1}^{K}\log\left(\frac{\Gamma(w_k)}{\Gamma(\widetilde{\alpha}_k)}\right) + \\ \sum_{k=1}^{K}(\widetilde{\alpha}_k - w_k)\cdot\left[\psi(\widetilde{\alpha}_k) - \psi\left(\sum_{k=1}^{K}\widetilde{\alpha}_k\right)\right] = \log\left(\frac{\Gamma\left(\sum_{k=1}^{K}\widetilde{\alpha}_k\right)\cdot\prod_{k=1}^{K}\Gamma\left(w_k\right)}{\Gamma\left(K\right)\cdot\prod_{k=1}^{K}\Gamma\left(\widetilde{\alpha}_k\right)}\right) + \\ + \sum_{k=1}^{K}(\widetilde{\alpha}_k - w_k)\cdot\left[\psi(\widetilde{\alpha}_k) - \psi\left(\sum_{k=1}^{K}\widetilde{\alpha}_k\right)\right] \end{aligned} \tag{5}$$

With Γ being the gamma function, and ψ being the digamma function.

3.2 Uncertainty-Aware Attention Buffer for Knowledge Distillation

Prompt optimization involves minimizing a knowledge distillation loss $\mathcal{L}_{KD}$ term between the model attention maps (computed through attention rollout) and the maps available in our *uncertainty-aware attention buffer* A shared within all the C clients of the federation, with each client providing M attention maps for each of the K classes:

$$A = \bigcup_{c=1}^{C}\bigcup_{k=1}^{K}\bigcup_{m=1}^{M} a_{c,k,m} \tag{6}$$

here $a_{c,k,i} \in \mathcal{R}^{H\times W}$ represents the i^{th} attention map for the k^{th} class of the c^{th} client of the federation. H and W are the height and width of the attention maps equal to input image dimensions. The **knowledge distillation loss** $\mathcal{L}_{KD}$ for a generic training sample of client c, with class k can be expressed as:

$$\mathcal{L}_{KD} = \frac{1}{M}\sum_{i=1}^{C\setminus c}\sum_{m=1}^{M}\left\|a_{c,k,\_} - a_{i,k,m}\right\|^2 \tag{7}$$

where, $\left\|a_{c,k,\_} - a_{i,k,m}\right\|^2$ denotes the squared Euclidean distance between the attention map $a_{c,k,\_}$ of the considered training sample and $a_{i,k,m}$ being an item of the buffer A.

The selection of samples for the attention buffer A by each client is based on the assumption that each local model should share its most confident predictions and indicate the image regions it focuses on. We use uncertainty scores from our evidential learning approach to guide the selection of attention maps for sharing within the federation. Specifically, we compute uncertainty scores, $u_j^{(k)}$, for each sample in class k, where j ranges from 1 to $N^{(k)}$, the total number of samples in class k. From these, we select the M samples with the lowest uncertainty scores, denoted as $\{u_1^{(k)}, u_2^{(k)}, \ldots, u_M^{(k)}\}$, and corresponding attention maps for inclusion in our uncertainty-aware attention buffer, A, replacing older ones.

4 Experimental Results

We validate the effectiveness of our proposed method on a multicenter dataset of 23,247 dermoscopic images of nine skin lesions from different populations and medical centers, based on the ISIC2019 dataset [2,3,18]. To carry out federated learning, we organized the dataset into six nodes, with each node representing data from a specific source:

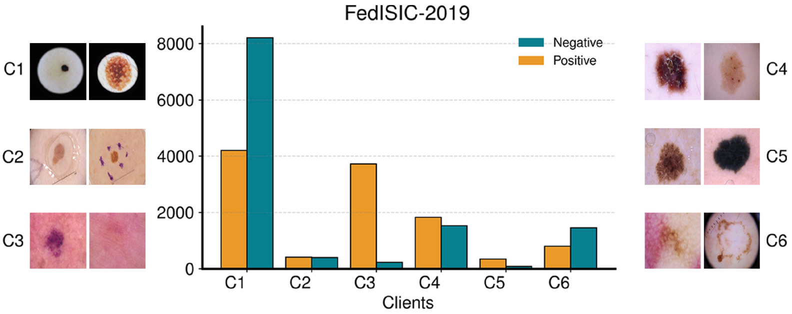


Fig. 2. Distribution of the Fed-ISIC2019 dataset across clients.

Client C1 contains the BCN20000 dataset as described by Combalia et al. [3], which includes 19,424 images from the Hospital Clínic in Barcelona; Clients C2, C3 and C4 are from the Austrian portion of the HAM10000 dataset [18], with images from the ViDIR Group at the Department of Dermatology at the Medical University of Vienna; Client C5 is also part of the HAM10000 dataset and contains the Rosendahl image set from the University of Queensland in Australia; while client C6 includes the MSK4 dataset [2]. The overall dataset exhibits heterogeneity in both the number of images contributed by each client as well ass in the distribution of classes, as illustrated in Fig. 2, making it a strong real-world use case for testing federated learning methods. For this study, we will focus on the binary classification task of distinguishing *Melanocytic nevus* from other skin lesions.

Training Procedure. In our setup, for each client, data is divided into a 75% training and 25% test split. Training is executed over 5 communication rounds, with 15 training epochs per round. Our model architecture employs a frozen ViT backbone augmented with additional parameters for b-prompts, t-prompts, and a classification-head. The ViT backbone specifications include an embedding dimension of 384, 6 attention heads, 12 blocks, and an input size of 224×224 pixels. For the b-prompts and t-prompts, the prompt keys (k and v) have a sequence length of 50, while l is 3 out of the 12 attention layers. We set the learning rates μ_1 and μ_2 to $2.5e-4$ and $5e-4$ respectively, with a weight decay factor of $1e-2$. Additionally, each client contributes 5 attention rollout maps per class (i.e., M in Eq. 6) to the uncertainty-aware attention buffer. Results are presented as in terms of balanced accuracy on the test set at the conclusion of all rounds.

Results. In Table 1, we present a comprehensive performance comparison between FedEvPrompt and existing federated learning methods. FedAvg [10] serves as our baseline. FedAvgPers builds upon FedAvg by integrating a personalization step through local data fine-tuning, aligning with our emphasis on personalized learning via b-prompts and t-prompts tuning. Additionally, we incorporate FedProx [8], specifically tailored to address non-IID data like our skin lesion dataset.

Given that our approach employs knowledge distillation without parameter sharing, we include two analogous methods in our analysis: FedProto [17] and FedDistill [15]. We also evaluate the performance of local training, where client models are trained independently without parameter sharing, using both sets of prompts (i.e., $(b,t)_{prompts}$), and using only one set of prompts ($g_{prompts}$ - general prompts) across all attention layers. We define *g-prompts* = [*b-prompts*, *t-prompts*] with both learning rates set to μ_1. This evaluation aims to validate our choice to apply different parameters across different attention layers and to demonstrate the advantages provided by our federated learning approach.

Results show that FedEvPrompt outperforms its competitors, including those that share parameters (thus being less privacy-preserving), such as FedAvg [10] and FedProx [8]. Notably, when comparing FedEvPrompt performance with other methods that do not share parameters, namely FedProto [17] and FedDistill [15], we observe higher performance across all clients and a lower standard deviation, indicating better convergence in accuracy among clients.

Table 1. Comparison with state-of-the-art methods on the Skin Lesion Dataset. In bold, best accuracy values.

	C1	**C2**	**C3**	**C4**	**C5**	**C6**	**Avg**
Local $b, t_{prompts}$	72.56	50.00	89.91	79.94	67.61	50.00	68.34±10.98
Local $g_{prompts}$	50.00	50.48	84.79	78.91	71.59	50.00	64.30±10.31
FedAvg [10]	74.74	72.79	67.74	79.89	67.61	77.09	73.31±3.64
FedAvgPers	77.79	68.01	84.84	78.08	67.61	82.66	76.50±7.25
FedProx [8]	81.34	68.14	74.69	75.15	77.84	78.65	75.97±4.55
FedProto [17]	71.88	69.04	66.34	69.71	58.52	73.57	68.18±5.34
FedDistill [15]	81.08	67.42	60.96	77.13	70.45	80.33	72.90±7.98
FedEvPrompt	81.02	71.83	84.15	79.02	68.18	79.35	**77.26**±4.65

We finally conducted an ablation study to assess the impact of various prompting options and sharing strategies among nodes within the federation. It's worth noting that while prompt sharing may potentially compromise privacy guarantees, exploring its effectiveness compared to using private prompts and our proposed knowledge distillation approach is interesting.

To this end, we initially assessed the performance of a variant of FedAvg where only low-level *b-prompts* are shared, gradually incorporating *t-prompts* and knowledge distillation on the uncertainty-aware attention buffer. Furthermore, we examined the variant of the proposed prompting strategy using a single set of general prompts *g-prompts* shared between nodes and coupled with our knowledge distillation method. Our findings, outlined in Table 2, underscore the significance of separate prompt learning, as evidenced by the subpar performance of the *g-prompts* variants. Interestingly, sharing separate sets of *b-prompts* and *t-prompts* (first two rows of Table 2) proved less effective than keeping them private and employing knowledge distillation (best performance observed in the last two rows of the same table). Moreover, we demonstrate that our strategy of incorporating attention maps based on uncertainty scores (as detailed in Sect. 3.2) yields superior performance compared to random selection of buffer samples. These two last considerations highlight the informative contribution provided by the attention maps corresponding to the lowest uncertainty samples driving clients' models towards the most significant regions of skin lesion images.

Table 2. Ablation study results showing the impact of shared prompts and knowledge distillation (KD) on federated learning performance.

	C1	**C2**	**C3**	**C4**	**C5**	**C6**	Avg
FedAvg $b_{prompts}$	50.00	70.04	70.29	79.61	71.59	79.21	70.18±7.56
+ $t_{prompts}$	74.74	72.79	67.74	79.89	67.61	77.09	73.31±3.64
+ KD	73.02	71.41	76.80	81.74	67.61	79.25	74.97±3.94
FedAvg $g_{prompts}$	50.00	70.88	72.77	50.00	67.61	50.00	60.21±6.81
FedAvg $g_{prompts}$ + KD	81.1	71.32	72.83	78.89	63.64	80.49	74.71±6.18
KD$_{random}$	78.12	69.40	76.80	80.30	68.75	77.65	75.17±3.66
KD$_{uncertainty}$ *(Ours)*	81.02	71.83	84.15	79.02	68.18	79.35	**77.26**±4.65

5 Conclusion

This work introduces FedEvPrompt, a new federated learning approach tailored for skin lesion classification using the ISIC2019 dataset. Indeed, this dataset offers a realistic setting for evaluating federated learning methods, eliminating the need for simulated distributions. FedEvPrompt seamlessly integrates evidential deep learning, prompt tuning, and knowledge distillation within a vision transformer architecture. Knowledge distillation on attention maps, in particular, ensures better privacy-preserving capabilities than parameter sharing. In addition to its superior performance in addressing data heterogeneity and privacy concerns, the employment of evidential learning offers enhanced model interpretability and uncertainty quantification, providing valuable insights for decision-making in medical image analysis. Balancing vacuity and dissonance [5,6] in buffer selection warrants further research to comprehensively understand underlying mechanisms.

Acknowledgements. We acknowledge the support of the PNRR ICSC National Research Centre for High Performance Computing, Big Data and Quantum Computing (CN00000013), under the NRRP MUR program funded by the NextGenerationEU. Rutger Hendrix is a PhD student enrolled in the National PhD in Artificial Intelligence, XXXVIII cycle BIS, course on Health and life sciences, organized by Università Campus Bio-Medico di Roma.

References

1. Abnar, S., Zuidema, W.: Quantifying attention flow in transformers. arXiv preprint arXiv:2005.00928 (2020)
2. Codella, N.C., et al.: Skin lesion analysis toward melanoma detection: a challenge at the 2017 international symposium on biomedical imaging (isbi), hosted by the international skin imaging collaboration (isic). In: 2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018), pp. 168–172. IEEE (2018)
3. Combalia, M., et al.: Bcn20000: Dermoscopic lesions in the wild (2019)
4. Geiping, J., Bauermeister, H., Dröge, H., Moeller, M.: Inverting gradients-how easy is it to break privacy in federated learning? Adv. Neural. Inf. Process. Syst. **33**, 16937–16947 (2020)
5. Guo, Z., et al.: A survey on uncertainty reasoning and quantification for decision making: Belief theory meets deep learning. arXiv preprint arXiv:2206.05675 (2022)
6. Josang, A., Cho, J.H., Chen, F.: Uncertainty characteristics of subjective opinions. In: 2018 21st International Conference on Information Fusion (FUSION), pp. 1998–2005. IEEE (2018)
7. Li, G., Wu, W., Sun, Y., Shen, L., Wu, B., Tao, D.: Visual prompt based personalized federated learning. arXiv preprint arXiv:2303.08678 (2023)
8. Li, T., Sahu, A.K., Zaheer, M., Sanjabi, M., Talwalkar, A., Smith, V.: Federated optimization in heterogeneous networks. Proc. Mach. Learn. Syst. **2**, 429–450 (2020)
9. Li, X.L., Liang, P.: Prefix-tuning: Optimizing continuous prompts for generation. arXiv preprint arXiv:2101.00190 (2021)
10. McMahan, B., et al.: Communication-efficient learning of deep networks from decentralized data. In: Artificial intelligence and statistics, pp. 1273–1282. PMLR (2017)
11. Pennisi, M., et al.: Feder: federated learning through experience replay and privacy-preserving data synthesis. Comput. Vis. Image Underst. **238**, 103882 (2024)
12. Proof:, J.: Kullback-leibler divergence for the dirichlet distribution, https://statproofbook.github.io/P/dir-kl.html, https://doi.org/10.5281/zenodo.4305949

13. Sensoy, M., Kaplan, L., Kandemir, M.: Evidential deep learning to quantify classification uncertainty. Adv. Neural Inform. Process. Syst. **31** (2018)
14. Sensoy, M., Kaplan, L., Kandemir, M.: Evidential deep learning to quantify classification uncertainty. In: Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N., Garnett, R. (eds.) Advances in Neural Information Processing Systems, vol. 31. Curran Associates, Inc. (2018), https://proceedings.neurips.cc/paper_files/paper/2018/file/a981f2b708044d6fb4a71a1463242520-Paper.pdf
15. Seo, H., Park, J., Oh, S., Bennis, M., Kim, S.L.: Federated knowledge distillation. Mach. Learn. Wireless Commun. 457 (2022)
16. Ståhl, N., Falkman, G., Karlsson, A., Mathiason, G.: Evaluation of uncertainty quantification in deep learning. In: Lesot, M.-J., et al. (eds.) IPMU 2020. CCIS, vol. 1237, pp. 556–568. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-50146-4_41
17. Tan, Y., et al.: Fedproto: federated prototype learning across heterogeneous clients. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 36, pp. 8432–8440 (2022)
18. Tschandl, P., Rosendahl, C., Kittler, H.: The ham10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. Sci. Data **5**(1), 1–9 (2018)
19. Wang, M., et al.: Federated uncertainty-aware aggregation for fundus diabetic retinopathy staging. arXiv preprint arXiv:2303.13033 (2023)
20. Zhu, L., Liu, Z., Han, S.: Deep leakage from gradients. Adv. Neural Inform. Process. Syst. **32** (2019)

Adaptive Text Feature Updating for Visual-Language Tracking

Xuexin Liu[1,2], Zhuojun Zou[1], and Jie Hao[1,3](✉)

[1] Institute of Automation, Chinese Academy of Sciences, Beijing, China
{liuxuexin2022,zouzhuojun2018,jie.hao}@ia.ac.cn
[2] School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing, China
[3] Guangdong Institute of Artificial Intelligence and Advanced Computing, Guangzhou, China

Abstract. Visual-language tracking combines visual and textual information to improve the accuracy of object tracking in video sequences. However, existing methods utilize language descriptions only from the initial frame of the video sequence, leading to inaccuracies as the target's appearance changes. To overcome this limitation, we propose a real-time updating method for language descriptions based on the target's current features. Our approach incorporates a large visual-language model that continuously generates descriptions to maintain relevance and accuracy, as well as an update determination module to assess whether the current text's quality requires refreshing. Additionally, we introduce a text fusion method to combine descriptions from different moments within the sequence, enhancing coherence and precision. Our method's effectiveness is validated on OTB99, LaSOT, and TNL2K datasets, demonstrating superior performance and adaptability in various tracking scenarios.

Keywords: Visual-Language Tracking · Large Visual-Languge Model · Online Feature Update

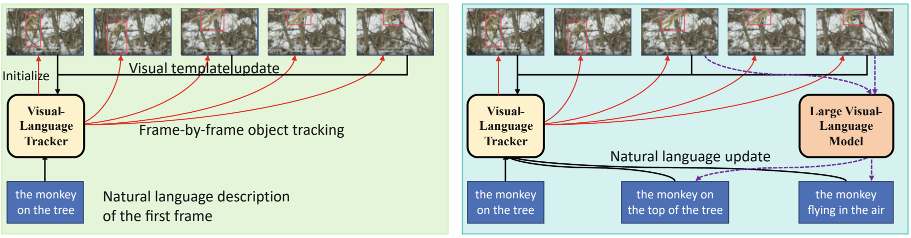
1 Introduction

Visual-Language Tracking (VLT), which integrates visual and textual information to locate targets in video sequences, has become a crucial task in computer vision [1]. Compared to conventional tracking approaches that rely solely on visual templates, VLT methods leverage language descriptions to guide the tracking process, offering improved adaptability and performance. Using text to describe and track objects can also enhance the effectiveness of systems in real-world applications like surveillance, autonomous driving, and human-computer interaction [2–4].

Current VLT methods [5–8] typically rely on a static textual description corresponding to the initial frame of the video sequence. While these approaches have achieved great results on datasets, they have inherent limitations. The main challenge is that the initial description does not adapt to changes in the

X. Liu and Z. Zou—Denotes equal contribution.

A. Antonacopoulos et al. (Eds.): ICPR 2024, LNCS 15329, pp. 366–381, 2025.
https://doi.org/10.1007/978-3-031-78110-0_24


Initialize
Visual template update
Visual-Language Tracker
Frame-by-frame object tracking
the monkey on the tree
Natural language description of the first frame
Visual-Language Tracker
Large Visual-Language Model
Natural language update
the monkey on the tree
the monkey on the top of the tree
the monkey flying in the air

(a) Update visual template during tracking

(b) Update visual template and natural language during tracking

Fig. 1. A comparison of existing image-updated methods and our proposed method. (a) Conventional methods only update the image template during tracking, overlooking that the correspondence between text and image changes with the target. (b) Our proposed method updates the text description along with the template, enhancing the stability of visual-language tracking.

object's appearance as the video progresses. Therefore, this static characteristic may lead to a decline in tracking accuracy over time. In contrast, in conventional visual tracking methods [9–13], image template updating techniques have already become relatively common. Typically, a series of updated templates of the target are stored and managed, as shown in Fig. 1(a). This kind of method is continuously updated online to accommodate changes in the target's appearance, thereby ensuring more accurate and robust tracking. Inspired by this approach and to address the issue of dynamic text-image mismatch in subsequent frames, we propose a similar strategy for VLT. Specifically, we introduce a method that updates the textual descriptions based on the target's current features. This online updating approach, illustrated in Fig. 1(b), utilizes a large visual-language model to generate and refresh the textual descriptions. By utilizing the multimodal understanding and generation capabilities of a large vision-language model, our method can effectively interpret and describe the visual features of the target, ensuring that the textual guidance remains relevant and accurate throughout the tracking process. Besides, we introduce an update determination module that determines whether the current text needs to be refreshed based on its quality. Additionally, we propose a text fusion method that integrates multiple descriptions from various moments to create a comprehensive representation of the target.

We validate the effectiveness of our text online updating method through comprehensive evaluations of three datasets with textual descriptions, including OTB99, LaSOT, and TNL2K. Furthermore, we test our approach across different initialization scenarios (NL, NLBBOX, BBOX). Our main contributions are as follows:

- We propose an online updating approach for VLT trackers, incorporating an innovative text fusion module and a text quality assessment mechanism. This approach dynamically updates descriptions to adapt to changes in the object's appearance, ensuring continuous and accurate tracking.

- We introduce a finetuned large vision-language model for updating tracker texts. This model leverages its multimodal understanding and generation capabilities to produce precise and context-aware descriptions, thereby enhancing tracking performance.
- Extensive experiments on three challenging datasets confirm the effectiveness of our proposed method, showcasing its robustness and adaptability across diverse tracking scenarios.

2 Related Work

2.1 Visual-Language Tracking

Given the abundance of semantic information in natural language, VLT has increasingly garnered the interest of researchers. TNLS [1] introduces the concept of "Tracking by Natural Language Specification(TNL)," aiming to harness the rich semantic information within the text to enhance human-computer interaction. GTI [14] breaks the TNL task into three distinct subtasks: Grounding, Tracking, and Integration. To overcome the lack of datasets in VLT, TNL2K [5] introduces a benchmark for object tracking with rich natural language descriptions. Despite significant progress in the aforementioned research, grounding and tracking methods remain separate and do not support end-to-end training. To address this issue, some studies design a unified framework by introducing a masking mechanism. For example, JointNLT [2] uses a Transformer module to fuse visual and textual cues, while UVLTrack [7] designs a multi-modal contrastive loss function, mapping visual and textual features to the same semantic space. Both approaches can utilize a single model to simultaneously complete grounding and tracking tasks, significantly enhancing usability.

Although previous methods improve tracker performance with language descriptions, they overlook the mismatch issue between text and the target's actual state. Textual cues are usually based on the initial frame, causing inconsistencies in later frames. Despite these issues, few studies focus on updating text features. QueryNLT [8] introduces a real-time adjustment module that selects features better suited to the current state by leveraging the relation between temporal visual templates and language descriptions. However, it only filters out mismatched words from the initial text without generating new, more precise descriptions. In this study, we use a large visual-language model to analyze the target in the current frame and generate accurate textual descriptions that match its state.

2.2 Large Visual-Launguage Model

In recent years, large language models (LLMs) [15–19] have achieved remarkable performance in a variety of language processing tasks due to their superior natural language understanding and interactivity. Expanding on the foundation provided by pre-trained LLMs, a series of visual-language models [20–24] have been developed. These methods have attained leading standards in numerous visual-linguistic tasks, including, but not limited to, referring expression generation, image captioning, visual reasoning, visual question answering, and visual grounding.

For instance, BLIP-2 [23] integrates image embeddings at the LLM's input stage and utilizes an auxiliary Q-Former to align the image and text features which are then concatenated and input into the LLM. Building upon the BLIP-2 framework, instructBLIP [24] replaces the descriptive text within the Q-Former with instructive text that contains querying information, enhancing the model's specificity for image-text tasks. Qwen-VL [20] introduces a visual encoder and a position-aware adapter into Qwen [17] and implements three phases of joint training with the LLM to specialize the pure language model for image-text tasks. CogVLM [21] augments the text attention model of LLMs with a parallel image model that has separate Query-Key-Value matrices and Feed-Forward Networks (FFNs) but collaborates in performing multi-head attention. While this approach considerably increases the size of the LLM, it maintains the language processing abilities and enriches the image comprehension proficiency of the large Vision Language Model (VLM), especially in terms of object localization.

Compared to conventional VLMs, LVLMs exhibit enhanced abilities, including (1) improved comprehension and expressive capabilities, (2) the ability to describe open-set targets, (3) versatility, and (4) interactivity. In this work, we employ the LoRA fine-tuning approach [25] to train a Large VLM for Referring Expression Generation on visual-language tracking datasets, thereby enabling the tracker to update descriptive text online based on the current image and target bounding box.

3 Methodology

The primary objective of this paper is to develop an effective plug-and-play method for visual-language tracking. Our goal is to develop an integrable solution that enhances existing models in generating referring expressions and feature fusion, thereby resolving issues with text-image mismatches. After outlining some fundamental concepts, this section will detail the methodology we employed to achieve this goal.

3.1 Preliminaries

In the context of visual-language tracking tasks, there are three sources of inputs: initial text (T_{1st}), template image (I_T), and search image (I_S). For the Tracking by Natural Language Specification (TNL) task, the inputs are T_{1st} and I_S. This task can be formulated as: TNL : $\{T_{1st}, I_S\} \rightarrow B$ where B represents the target bounding box. Initially, T_{1st} is used to locate B, and then object tracking is performed. Conversely, the Tracking by Language and Box Specification (TNLBBOX) task utilizes T_{1st}, I_T, and I_S, which can be expressed as

TNLBBOX : $\{T_{1st}, I_T, I_S\} \rightarrow B$. Here, T_{1st} is used to enhance tracking performance, given an initial bounding box B_0. Additionally, visual object tracking, also known as Tracking by Box Specification (TBBOX), can be defined as: TBBOX : $\{I_T, I_S\} \rightarrow B$.

Current research trends favor using a unified network framework to handle all tasks by employing masking techniques. Specifically, in the NL task, the I_S is replaced with a zero-padded template. Subsequently, the visual and textual information are processed through feature encoders to obtain the corresponding visual and textual features. These features are then concatenated for feature fusion. After fusion, the features of the search region are extracted and passed through a prediction head or decoder to determine the position of the identified object bounding box. In our research, we propose an innovative update method, which includes not only the initial text T_{1st} but also the updated text T_{upd} generated by a large visual-language model based on the current frame state, as shown in Fig. 2. This design aims to better align textual features with the current target appearance. By ensuring that the textual information remains relevant and accurately reflects the evolving context, we can enhance the accuracy and robustness of target tracking.

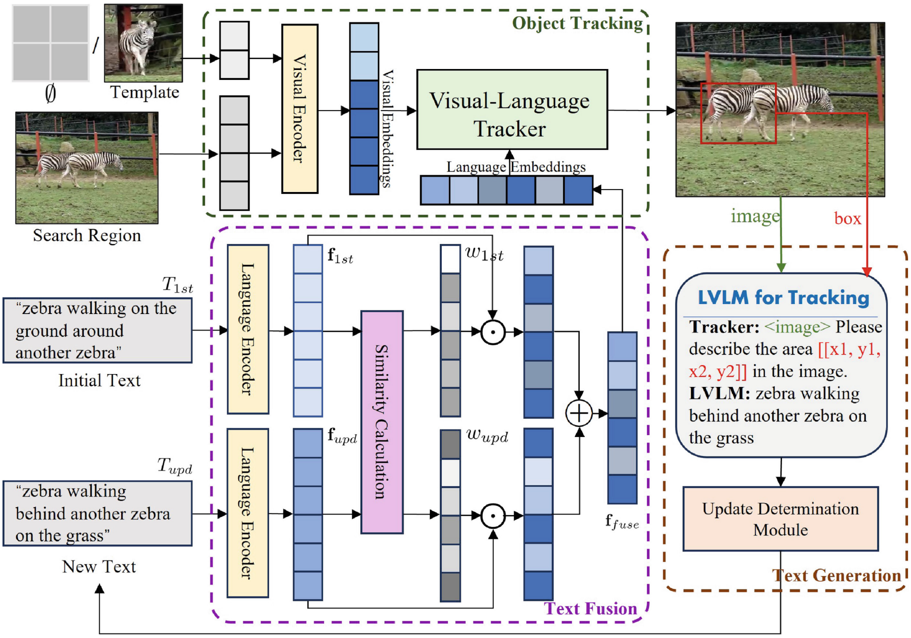


Fig. 2. The pipeline of our text update process. The model is divided into three parts: object tracking, text generation, and text fusion. The results estimated by the object tracker, along with the current image, are given to the LVLM to generate new text. After being evaluated by the update determination module, the generated text is fused with the initial text to create new text embeddings, supporting subsequent tracking.

3.2 LVLM-Based Referring Expression Generation

During tracking, generating descriptions based on the target's current visual features aligns with the goal of the Referring Expression Generation (REG) task, which is to describe objects within a specific region of an image. However, a core issue with present REG methods, including those LVLM models when applied to tracking sequences, is the poor distinctiveness of the generated descriptions. When there are similar objects in the image, the visual-language tracker can mistakenly localize to the wrong object due to ambiguous descriptions (see Sect. 4.4 for more details). To meet the requirements of the object tracking task, we select CogVLM as the foundational referring expression generator and fine-tune it to obtain CogVLM_Track, which is adapted for the tracking scenario.

As illustrated in Fig. 2, our method takes the current image and target location as inputs and outputs a description of the target. Unlike general scenarios that require LVLM to have robust interactive capabilities, the tracking task only carries out user-invisible queries on LVLM. Therefore, we only use a single prompt during training and testing, which is:

$$pmt(box) = [pmt_{pre}, [x_1, y_1, x_2, y_2], pmt_{suf}], \tag{1}$$

where x_1, y_1, x_2, y_2 represent the horizontal and vertical coordinates of the top-left and bottom-right corners of the target. Each coordinate value is discretized to an integer within 1000 based on the original image size and is padded with leading zeros to three digits as proposed in [21]. pmt_{pre} and pmt_{suf} are the fixed prefix and suffix of the prompt, respectively[1].

To generate textual descriptions of the target, we train the network using cross-entropy loss:

$$\mathcal{L}_{REG}(\theta) = -\log P(w_t | w_1, w_2, \ldots, w_{t-1}, I, pmt(box)), \tag{2}$$

where θ represents the trainable parameters, w_t is the predicted word at position t, and I is the current image.

To filter out invalid descriptions in advance, we use an update determination module to assess the quality of the generated text. The formula is as follows:

$$\begin{gathered} T_{upd} = \begin{cases} T_{new} & if \ \mathrm{IoU}(box, \hat{box}) > \sigma, \\ \tilde{T}_{upd} & elsewise. \end{cases} \\ s.t. \ T_{new} = LVLM(I, box), \ \hat{box} = VG(I, T_{new}). \end{gathered} \tag{3}$$

In this module, we use the text T_{new} generated by LVLM and employ the visual-language tracker's visual grounding function $VG(\cdot)$ to relocate the target in the image, yielding $\hat{\text{box}}$. We then determine the quality of T_{new} by checking if the IoU between $\hat{\text{box}}$ and the original box exceeds the threshold σ, deciding whether to update T_{upd} with T_{new} or maintain the history text $\hat{T}_{\text{upd}}$.

[1] pmt_{pre}: *Please describe the area;*
pmt_{suf}: *in the image, including color or positional information for distinction.*

3.3 Online Text Feature Fusion

The current mainstream trackers construct the image template by preserving the initial frame and periodically updating subsequent frames. To adapt to the feature composition on the image side, we do not directly use the text embeddings generated by LVLM. Instead, they are fused with the initial description before being fed into the tracker to ensure that the final feature representation maintains both cross-modal temporal consistency and contextual accuracy.

As shown in Fig. 2, we first use a text encoder to perform embeddings on texts T_{1st} and T_{upd}, obtaining feature vectors $\mathbf{f}_{1st}$ and $\mathbf{f}_{upd}$, respectively. It should be noted that since T_{1st} remains constant during tracking, $\mathbf{f}_{1st}$ is computed only once at initialization and is reused thereafter. Subsequently, a cosine similarity function is applied to generate the attention weights w_{upd}, with the formula as follows:

$$w_{upd} = \cos\_\mathrm{sim}(\mathbf{f}_{1st}, \mathbf{f}_{upd}) = \frac{\mathbf{f}_{1st} \cdot \mathbf{f}_{upd}}{\|\mathbf{f}_{1st}\| \|\mathbf{f}_{upd}\|}, \tag{4}$$

where $\cdot$ represents the dot product, and $\| \cdot \|$ represents the Euclidean norm of a vector. We use the weighted sum $\mathbf{f}_{fuse}$ of $\mathbf{f}_{1st}$ and $\mathbf{f}_{upd}$ as the tracker's input, setting the w_{1st} to $1 - w_{upd}$ to ensure $\sum w = 1$, which is represented as follows:

$$\mathbf{f}_{fuse} = (1 - w_{upd}) \cdot \mathbf{f}_{1st} + w_{upd} \cdot \mathbf{f}_{upd}. \tag{5}$$

During tracking, $\mathbf{f}_{1st}$ initializes $\mathbf{f}_{upd}$. When in the first frame or without a text generator, the text fusion module degenerates to output the initial features $\mathbf{f}_{1st}$. This property ensures the robustness of the method.

Additionally, to facilitate the subsequent processing of text features, we provide a text mask $\mathbf{M}_{fuse}$ for $\mathbf{f}_{fuse}$ that is the union of the masks for $\mathbf{f}_{1st}$ initializes $\mathbf{f}_{upd}$, making all available features visible:

$$\mathbf{M}_{fuse}[i] = \mathbf{M}_{1st}[i] \vee \mathbf{M}_{upd}[i], \quad s.t. \forall i \in \{1, \dots, n\}. \tag{6}$$

where n is the length of the text embedding, $\mathbf{M}_{1st}$ and $\mathbf{M}_{upd}$ are the masks for $\mathbf{f}_{1st}$ and $\mathbf{f}_{upd}$ respectively, and $\vee$ represents the bitwise OR operation.

By employing the above feature fusion method, we can ensure that the most relevant features from both the initial and updated texts are combined in a way that reflects their contextual importance.

4 Experiment

4.1 Implementation Details

LVLM for Tracking. During the training of CogVLM-Track, we utilize the first frame of the videos in training splits from OTB99 [26], LaSOT [27], and TNL2K [5] to create our referring expression generation dataset. Given the tendency of LaSOT's text to provide uniform descriptions for targets within the same object class, we select only one random video per class to include in our training data.

We set the batch size to 4 for each GPU, set the learning rate to 1e-5, and employ an Adam [28] optimizer with a weight decay of 0.05 to conduct the fine-tuning process over 4000 iterations on eight A800 GPUs.

The network architecture involved in fine-tuning includes the 63 layers of Transformer from ViT [29] used for image feature extraction and the final 32 layers of Transformer in CogVLM [21]. LoRA branches are configured for the respective QKV (Query, Key, Value) matrices and linear layers. During fine-tuning, the rank is set to 10 to ensure a minimal increase in the model size and rapid training. The original model size is 32.857 GB, and after fine-tuning, the model size is 32.893 GB, with an increase of only 0.1%.

Online Text Update. To align the textual description with the tracking image, we invoke LVLM to generate the description using the cropped search region of the current frame together with the corresponding target bounding box, rather than using the full image. The parameter σ utilized by the update determination module is set to 0.5 across all sequences. In our online update setting, the text updates are scheduled to coincide with the image template updates. The baseline visual-language tracker we use is UVLTrack [7], for which we conducted NL, NLBBOX, and BBOX Tracking experiments mainly on its *large* version. It should be noted that the results we provide show slight deviations from those provided by [7], which is due to our use of the official code and official network weights retested in our environment to ensure a fair comparison.

Datasets and Metrics. We evaluate our method on three tracking benchmark datasets with natural language descriptions, which include OTB99 [26], LaSOT [27], and TNL2K [5]. The OTB99 dataset, with textual descriptions provided by Li et al., contains 99 sequences, 48 of which are designated for testing. The LaSOT dataset is a long-term tracking dataset that provides bounding boxes and natural language descriptions, featuring 280 long-term test sequences. The TNL2K dataset, specifically designed for vision language tracking tasks, includes 700 test sequences covering both real and synthetic scenes such as cartoon videos, and it presents challenging factors like significant appearance changes and adversarial samples. In this study, we employ three key metrics to evaluate the performance of trackers: Area Under Curve (AUC) of Success Plot, Precision (P), and Normalized Precision (P_N). AUC measures overall performance under various conditions; Precision reflects the degree of overlap between the tracking results and the actual targets; Normalized Precision adjusts the precision metric considering changes in target size.

4.2 Evaluations on 3 Tasks

We apply the proposed online text update method to three different tracking tasks to verify the generality of our approach.

NL Tracking. This task serves as the primary scenario for our verification. We conducted experiments on both the base and large versions of UVLTrack. From the top part of Table 1, it can be observed that applying our update method

Table 1. Comparison of our method with state-of-the-art trackers on OTB99 [26], LaSOT [27], and TNL2K [5] datasets. We highlight the top-ranked scores in red.

Method	OTB99			LaSOT			TNL2K		
	AUC	P	P$_N$	AUC	P	P$_N$	AUC	P	P$_N$
NL									
TNLS-II [1]	25.0	29.0	–	–	–	–	–	–	–
GTI [14]	58.1	73.2	–	47.8	47.6	–	–	–	–
TNL2K-1 [5]	19.0	24.0	–	51.1	49.3	–	11.4	6.4	11.0
CTRNLT [30]	53.0	72.0	–	52.0	51.0	–	14.0	9.0	–
JointNLT [2]	59.2	77.6	–	56.9	59.3	64.5	54.6	55.0	70.6
QueryNLT [8]	61.2	81.0	73.9	54.2	55.0	62.5	53.3	53.0	70.4
UVLTrack-B* [7]	60.1	77.0	71.1	56.7	60.2	64.2	54.7	56.3	70.9
UVLTrack-L* [7]	62.6	81.5	73.7	59.3	63.5	66.8	58.2	60.8	74.6
UVLTrack-B* + ours	60.9	78.1	72.1	57.1	60.8	64.6	54.9	56.7	71.2
UVLTrack-L* + ours	63.1	82.1	74.3	59.7	63.9	67.3	58.3	60.8	74.7
NLBBOX									
TNLS-III [1]	55.0	72.0	–	–	–	–	–	–	–
SNLT [31]	66.6	80.4	–	54.0	57.6	-	27.6	41.9	–
TNL2K-2 [5]	68.0	88.0	–	51.0	55.0	-	41.7	42.0	50.0
JointNLT [2]	65.3	85.6	79.5	60.4	63.6	69.4	56.9	58.1	73.6
QueryNLT [8]	66.7	88.2	82.4	59.9	63.5	69.6	57.8	58.7	75.6
UVLTrack-L* [7]	70.8	92.2	86.3	70.8	78.0	81.1	64.7	69.0	82.5
UVLTrack-L*+ours	70.9	92.5	86.7	71.0	78.2	81.3	65.1	69.4	82.9
BBOX									
Ocean [32]	–	–	–	56.0	56.6	65.1	38.4	37.7	45.0
TransT [33]	–	–	–	64.9	69.0	73.8	50.7	51.7	–
MixFormer-L [34]	–	–	–	70.1	76.3	79.9	–	–	–
SimTrack-L [35]	–	–	–	70.5	-	79.7	55.6	55.7	–
UVLTrack-L* [7]	70.1	91.0	84.7	70.8	77.7	81.0	64.9	69.1	82.7
UVLTrack-L* +ours	70.7	92.0	85.8	71.0	78.0	81.3	65.1	69.4	83.0

* our reproducing results using the officially released code.

leads to improvements for both UVLTrack-B and UVLTrack-L, thereby asserting the effectiveness of our approach. Specifically, the proposed method achieved AUC scores of 63.1, 59.7, and 58.3 on OTB99, LaSOT, and TNL2K respectively, attaining leading performance. We noticed that the improvement of our method over the baseline is not substantial. This is because most trackers use a low update frequency (e.g., every 50 frames) for maintaining optimal tracking speed. To demonstrate the versatility of our method, we didn't artificially increase text update frequency on UVLTrack but allowed it to update concurrently with the

Table 2. Analysis of different components in our methods.

Method	UVLTrack_base (NL)								
	OTB99			LaSOT			TNL2K		
	AUC	P	P_N	AUC	P	P_N	AUC	P	P_N
baseline	60.1	77.0	71.1	56.7	60.2	64.2	54.7	56.3	70.9
+TUM(cos)	60.7	77.9	71.9	56.8	60.4	64.3	54.8	56.6	71.1
+TUM(avg)+UDM	60.8	77.9	71.9	57.0	60.6	64.6	54.8	56.6	71.1
+TUM(cos)+UDM	**60.9**	**78.1**	**72.1**	**57.1**	**60.8**	**64.6**	**54.9**	**56.7**	**71.2**

image template. As a consequence, our method had limited activations during tracking, which decreased opportunities to improve upon the base tracker.

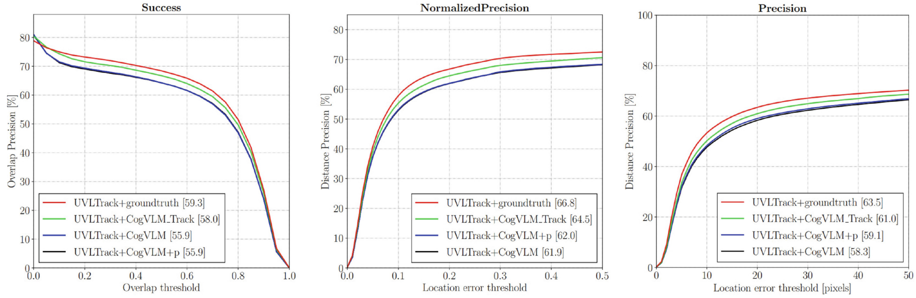
NLBBOX Tracking. In this task, UVLTrack is initialized with both the first frame image and text, and subsequently updated by our method. As shown in the middle part of Table 1, UVLTrack_L+ours outperforms the baseline across all metrics in three datasets. This demonstrates that even in tasks where tracking is predominantly image template-driven, our method can still provide richer scene information to the base tracker, enhancing tracking performance comprehensively.

BBOX Tracking. Although this task does not provide additional text annotations for tracking initialization, the natural language we provide does not depend on initial annotations like QueryNLT. Therefore, we can still perform online text generation to assist tracking. The bottom part of Table 1 shows an overall improvement of our method over the baseline, proving the generality of our approach in pure visual tracking Scenarios.

4.3 Ablation Study and Further Analysis

Ablation Study on Different Components. We ensure that LVLM is available in this experiment, as it serves as the foundation for text updating. Under this premise, we primarily verify the importance of two components: the Text Update Module (TUM) and the Update Determination Module (UDM). Among them, TUM corresponds to the Text Fusion part shown in Fig. 2. Table 2 presents the outcomes of applying TUM directly without UDM for text filtering (line #2). It is observable that there is an improvement in performance across all datasets relative to the baseline (line #1), demonstrating the efficacy of TUM. By incorporating UDM into TUM (line #4), the tracker's performance is further enhanced, highlighting the importance of the UDM.

Different Similarity Measurement Methods. We use averaging as a substitute for Eq. 5 to compare our approach with a simpler fusion method, formally expressed as $\mathbf{f}_{fuse} = 0.5 * \mathbf{f}_{1st} + 0.5 * \mathbf{f}_{upd}$. As shown in line #3 in Table 2, using this method shows improvement over the baseline method (line #1), indicating the high quality of our text generation which can be paired with different fusion methods. Additionally, the performance of this method across all datasets is lower than our approach (line #4), which validates the correctness of our fusion method.



(a) Success Plot (b) Normalized Precision Plot (c) Precision Plot

Fig. 3. Tracking performance comparison on the LaSOT dataset using texts generated by different models to replace the ground truth of the first frame in UVLTrack-large.

Different Text Generation Models. To assess text generation quality across models, we substituted the initial frame's ground truth with text from various models and compared their tracking performance. The models evaluated are CogVLM_Track (fine-tuned), CogVLM (original), and CogVLM + p (CogVLM with prompt). Results in Fig. 3 show that CogVLM_Track's tracking performance, with a 58.0 AUC on LaSOT, closely approaches ground truth performance, surpassing CogVLM and CogVLM+p by 2.1 points. Similarly, The superiority of CogVLM_Track on the P and P_N metrics indicates its generated text's high fidelity. Although slightly underperforming compared to ground truth, CogVLM_Track's text remains more accurate over time than maintaining the first frame description, suggesting the benefits of periodic updates (refer to Fig. 1 and Fig. 4 for comparison).

Inference Speed Analysis. In experiments using a single A100 GPU (40G), we test inference on image sequences with a resolution of 1280 × 720. The baseline inference speed of UVLTrack-B was 35 FPS, and with our text updating method, it slightly decreased to 33 FPS. This indicates that while inference speed slightly decreases, our method's dynamic adaptation to new visual features improves accuracy and robustness, particularly when text descriptions do not match visual information, making it valuable for complex environments.

4.4 Visualization

Generated Text. In the tracking task, language descriptions must accurately and distinctively describe the target to avoid confusion with similar objects.

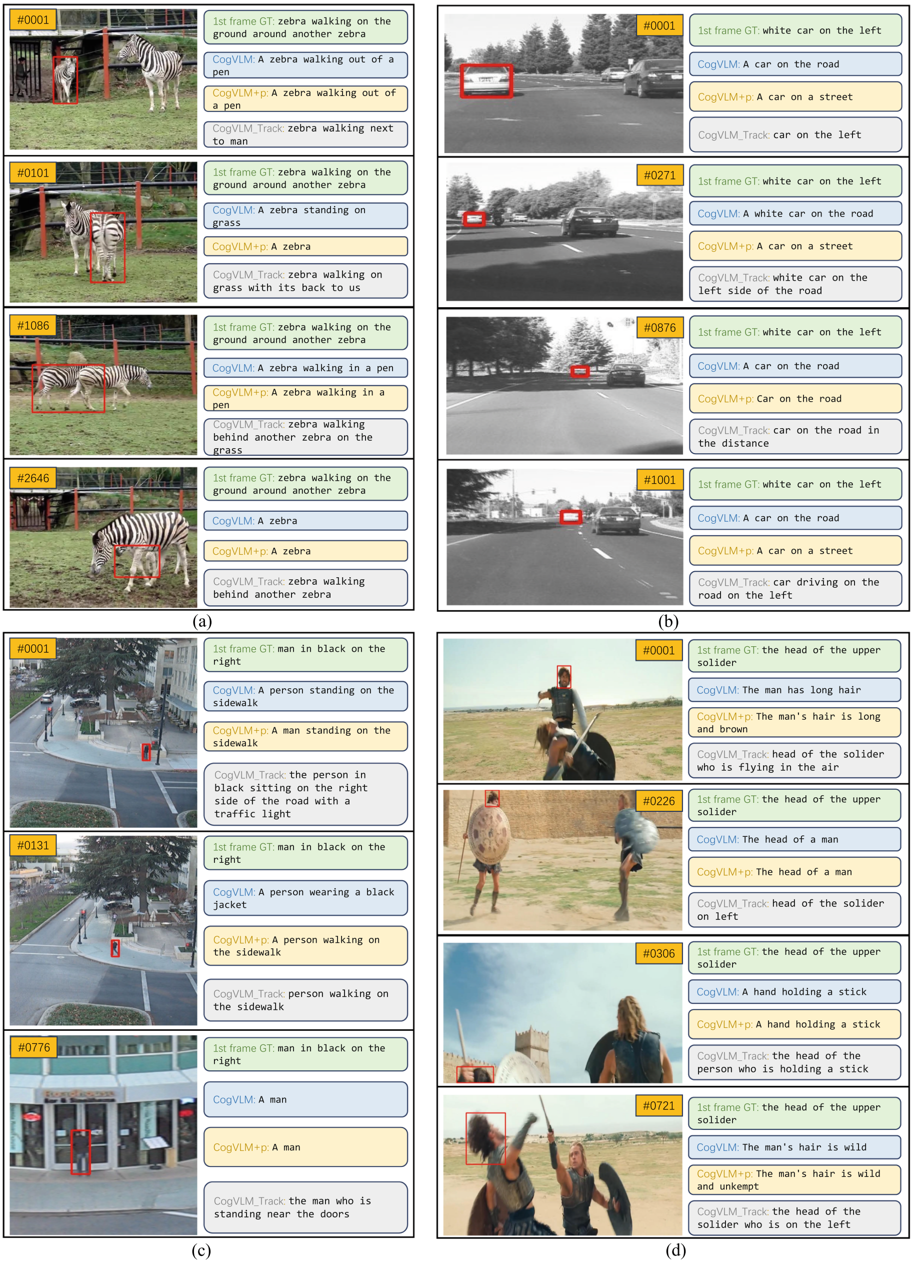

#0001
1st frame GT: zebra walking on the ground around another zebra
CogVLM: A zebra walking out of a pen
CogVLM+p: A zebra walking out of a pen
CogVLM_Track: zebra walking next to man
#0101
1st frame GT: zebra walking on the ground around another zebra
CogVLM: A zebra standing on grass
CogVLM+p: A zebra
CogVLM_Track: zebra walking on grass with its back to us
#1086
1st frame GT: zebra walking on the ground around another zebra
CogVLM: A zebra walking in a pen
CogVLM+p: A zebra walking in a pen
CogVLM_Track: zebra walking behind another zebra on the grass
#2646
1st frame GT: zebra walking on the ground around another zebra
CogVLM: A zebra
CogVLM+p: A zebra
CogVLM_Track: zebra walking behind another zebra
(a)
#0001
1st frame GT: white car on the left
CogVLM: A car on the road
CogVLM+p: A car on a street
CogVLM_Track: car on the left
#0271
1st frame GT: white car on the left
CogVLM: A white car on the road
CogVLM+p: A car on a street
CogVLM_Track: white car on the left side of the road
#0876
1st frame GT: white car on the left
CogVLM: A car on the road
CogVLM+p: Car on the road
CogVLM_Track: car on the road in the distance
#1001
1st frame GT: white car on the left
CogVLM: A car on the road
CogVLM+p: A car on a street
CogVLM_Track: car driving on the road on the left
(b)
#0001
1st frame GT: man in black on the right
CogVLM: A person standing on the sidewalk
CogVLM+p: A man standing on the sidewalk
CogVLM_Track: the person in black sitting on the right side of the road with a traffic light
#0131
1st frame GT: man in black on the right
CogVLM: A person wearing a black jacket
CogVLM+p: A person walking on the sidewalk
CogVLM_Track: person walking on the sidewalk
#0776
1st frame GT: man in black on the right
CogVLM: A man
CogVLM+p: A man
CogVLM_Track: the man who is standing near the doors
(c)
#0001
1st frame GT: the head of the upper solider
CogVLM: The man has long hair
CogVLM+p: The man's hair is long and brown
CogVLM_Track: head of the solider who is flying in the air
#0226
1st frame GT: the head of the upper solider
CogVLM: The head of a man
CogVLM+p: The head of a man
CogVLM_Track: head of the solider on left
#0306
1st frame GT: the head of the upper solider
CogVLM: A hand holding a stick
CogVLM+p: A hand holding a stick
CogVLM_Track: the head of the person who is holding a stick
#0721
1st frame GT: the head of the upper solider
CogVLM: The man's hair is wild
CogVLM+p: The man's hair is wild and unkempt
CogVLM_Track: the head of the solider who is on the left
(d)

Fig. 4. Visualization analysis of generated texts from CogVLM, CogVLM+p, and CogVLM_Track. The sequences are from the test sets of (a) LaSOT, (b)(c) OTB99, and (d) TNL2K, with the bounding boxes as annotations. We also show the ground truth of the first frame for comparison.

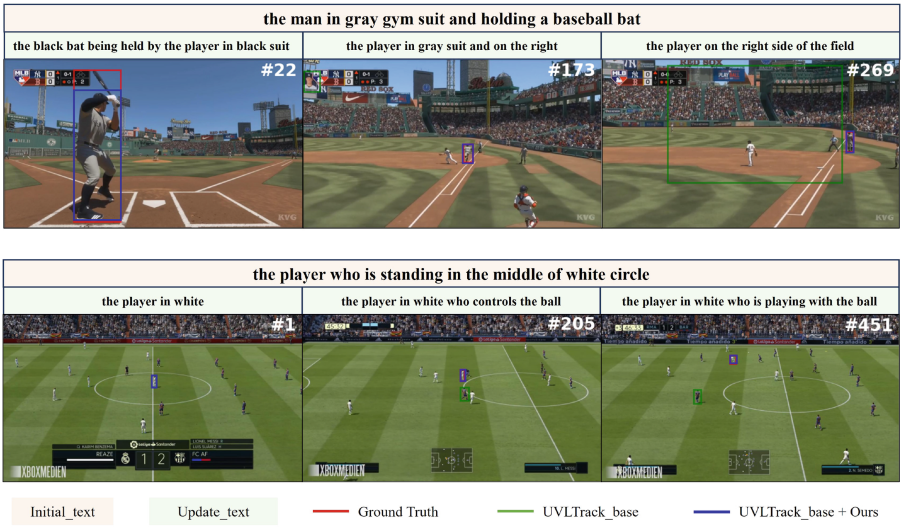

the man in gray gym suit and holding a baseball bat
the black bat being held by the player in black suit
the player in gray suit and on the right
the player on the right side of the field
#22
#173
#269
the player who is standing in the middle of white circle
the player in white
the player in white who controls the ball
the player in white who is playing with the ball
#1
#205
#451
Initial_text
Update_text
Ground Truth
UVLTrack_base
UVLTrack_base + Ours

Fig. 5. Visualization analysis of the effects of text updates on challenging video sequences from the TNL2K test set. (Zoom in for a better view).

From Fig. 4(a) and (b), it's evident that determining the target from CogVLM and CogVLM+p descriptions, and even from the first frame ground truth of Fig. 4(a), is challenging due to their lack of distinctiveness. In contrast, CogVLM-_Track provides specific instance descriptions. Additionally, Fig. 4(c) and (d) show that without changing the initial text, it's difficult to accurately describe the subsequent states, highlighting the necessity of text updating.

Text Update Validity. As shown in Fig. 5, we present a comparative analysis of two challenging sequences in TNL2K. The figure includes the initial text, the texts generated at the early, middle, and late stages of the tracking process, as well as the location box. The results demonstrate that because the generated texts better match the current state of the object, our method can locate the object more accurately compared to the baseline. For example, in frame #173 of the first sequence in the figure, the generated text "`the player in a gray suit and on the right`" provides better directional information and introduces new details that were not present in the initial description, helping the tracker to locate the object more precisely.

5 Conclusion

In this paper, we introduce a novel online text updating method for Visual-Language Tracking. We leverage the natural language processing capabilities of LVLM to design a text generation module and introduce a complementary text fusion module to maintain cross-modal temporal consistency with image templates. This method dynamically updates the entire textual description based on the target's current appearance, ensuring continuous and accurate tracking. Our approach outperforms the baseline approach in 3 different tracking tasks on the OTB99, LaSOT, and TNL2K datasets, demonstrating its superiority. We hope that our work will facilitate the further application of large models in the field of object tracking and the exploration of more efficient online text updating approaches.

Acknowledgement. This work was supported by the National Science and Technology Major Project from Ministry of Science and Technology, China (Grant No. 2018AAA0103100) and the National Natural Science Foundation of China under Grant 62236007.

References

1. Li, Z., Tao, R., Gavves, E., Snoek, C.G.M., Smeulders, A.W.M.: Tracking by natural language specification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 6495–6503 (2017)
2. Zhou, L., Zhou, Z., Mao, K., He, Z.: Joint visual grounding and tracking with natural language specification. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 23151–23160 (2023)
3. Zhao, X., Zhang, D., Liyuan, H., Zhang, T., Bo, X.: Ode-based recurrent model-free reinforcement learning for pomdps. Adv. Neural. Inf. Process. Syst. **36**, 65801–65817 (2023)
4. Wang, Z., Zhu, X., Zhang, T., Wang, B., Lei, Z.: 3d face reconstruction with the geometric guidance of facial part segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 1672–1682 (2024)
5. Wang, X., et al.: Towards more flexible and accurate object tracking with natural language: Algorithms and benchmark. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 13763–13773 (2021)
6. Zhang, C., et al.: All in one: exploring unified vision-language tracking with multi-modal alignment. In: Proceedings of the 31st ACM International Conference on Multimedia, pp. 5552–5561 (2023)
7. Ma, Y., Tang, Y., Yang, W., Zhang, T., Zhang, J., Kang, M.: Unifying visual and vision-language tracking via contrastive learning. arXiv preprint arXiv: 2401.11228 (2024)
8. Shao, Y., He, S., Ye, Q., Feng, Y., Luo, W., Chen, J.: Context-aware integration of language and visual references for natural language tracking. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 19208–19217 (2024)
9. Yan, B., Peng, H., Fu, J., Wang, D., Lu, H.: Learning spatio-temporal transformer for visual tracking. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 10448–10457 (2021)
10. Ye, B., Chang, H., Ma, B., Shan, S., Chen, X.: Joint feature learning and relation modeling for tracking: a one-stream framework. In: Avidan, S., Brostow, G., Cissé, M., Farinella, G.M., Hassner, T. (eds.) Computer Vision – ECCV 2022. ECCV 2022. LNCS, vol. 13682. Springer, Cham (2022). https://doi.org/10.1007/978-3-031-20047-2_20

11. Chen, X., Peng, H., Wang, D., Lu, H., Hu, H.: Seqtrack: sequence to sequence learning for visual object tracking. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 14572–14581 (2023)
12. Wei, X., Bai, Y., Zheng, Y., Shi, D., Gong, Y.: Autoregressive visual tracking. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 9697–9706 (2023)
13. Zou, Z., Hao, J., Shu, L.: Online feature classification and clustering for transformer-based visual tracker. In: 2022 26th International Conference on Pattern Recognition (ICPR), pp. 3514–3521. IEEE (2022)
14. Yang, Z., Kumar, T., Chen, T., Jingsong, S., Luo, J.: Grounding-tracking-integration. IEEE Trans. Circuits Syst. Video Technol. **31**(9), 3433–3443 (2020)
15. Achiam, J., et al.: Gpt-4 technical report. arXiv preprint, arXiv: 2303.08774 (2023)
16. Anil, R., et al.: Palm 2 technical report. arXiv preprint, arXiv: 2305.1040 (2023)
17. Bai, J., et al.: Qwen technical report. arXiv preprint, arXiv: 2309.16609 (2023)
18. Brown, T., et al.: Language models are few-shot learners. Adv. neural inform. process. syst. **33**, 1877–1901 (2020)
19. Zhang, R., et al.: Llama-adapter: Efficient fine-tuning of language models with zero-init attention. arXiv preprint, arXiv: 2303.16199 (2023)
20. Bai, J., et al.: Qwen-vl: A frontier large vision-language model with versatile abilities. arXiv preprint, arXiv: 2308.12966 (2023)
21. Wang, W., et al.: Cogvlm: Visual expert for pretrained language models. arXiv preprint, arXiv: 2311.03079 (2023)
22. Hong, W., et al.: Cogagent: a visual language model for gui agents. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 14281–14290 (2024)
23. Li, J., Li, D., Savarese, S., Hoi, S.: Blip-2: bootstrapping language-image pre-training with frozen image encoders and large language models. In: International Conference on Machine Learning, pp. 19730–19742. PMLR (2023)
24. Dai, W., et al.: Instructblip: towards general-purpose vision-language models with instruction tuning. Adv. Neural Inform. Process. Syst. **36** (2024)
25. Hu, E.J., et al.: Lora: Low-rank adaptation of large language models arXiv preprint arXiv: 2106.09685 (2021)
26. Li, Z., Tao, R., Gavves, E., Snoek, C.G.M., Smeulders, A.W.M.: Tracking by natural language specification. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 7350–7358 (2017)
27. Fan, H., et al.: Lasot: ahigh-quality benchmark for large-scale single object tracking. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 5374–5383 (2019)
28. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint, arXiv: 1412.6980 (2014)
29. Dosovitskiy, A.: An image is worth 16x16 words: transformers for image recognition at scale. In: International Conference on Learning Representations (2021)
30. Li, Y., Yu, J., Cai, Z., Pan, Y.: Cross-modal target retrieval for tracking by natural language. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 4931–4940 (2022)
31. Feng, Q., Ablavsky, V., Bai, Q., Sclaroff, S.: Siamese natural language tracker: Tracking by natural language descriptions with siamese trackers. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 5851–5860 (2021)

32. Zhang, Z., Peng, H., Fu, J., Li, B., Hu, W.: Ocean: object-aware anchor-free tracking. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J.-M. (eds.) ECCV 2020. LNCS, vol. 12366, pp. 771–787. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-58589-1_46
33. Chen, X., Yan, B., Zhu, J., Wang, D., Yang, X., Lu, H.: Transformer tracking. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 8126–8135 (2021)
34. Cui, Y., Jiang, C., Wang, L., Wu, G.: Mixformer: end-to-end tracking with iterative mixed attention. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 13608–13618 (2022)
35. Chen, B., et al.: Backbone is all your need: a simplified architecture for visual object tracking. In: Avidan, S., Brostow, G., Cissé, M., Farinella, G.M., Hassner, T. (eds.) Computer Vision – ECCV 2022, pp. 375–392, Springer Nature Switzerland, Cham (2022). https://doi.org/10.1007/978-3-031-20047-2_22

TrajDiffuse: A Conditional Diffusion Model for Environment-Aware Trajectory Prediction

Qingze Tony Liu[1] (✉), Danrui Li[1], Samuel S. Sohn[1], Sejong Yoon[2], Mubbasir Kapadia[1], and Vladimir Pavlovic[1]

[1] Rutgers University, Piscataway, USA
{tony.liu,danrui.li,mubbasir.kapadia}@rutgers.edu,
{sss286,vladimir}@cs.rutgers.edu
[2] The College of New Jersey, Ewing, USA
yoons@tcnj.edu

Abstract. Accurate prediction of human or vehicle trajectories with good diversity that captures their stochastic nature is an essential task for many applications. However, many trajectory prediction models produce unreasonable trajectory samples that focus on improving diversity or accuracy while neglecting other key requirements, such as collision avoidance with the surrounding environment. In this work, we propose `TrajDiffuse`, a planning-based trajectory prediction method using a novel guided conditional diffusion model. We form the trajectory prediction problem as a denoising impaint task and design a map-based guidance term for the diffusion process. `TrajDiffuse` is able to generate trajectory predictions that match or exceed the accuracy and diversity of the SOTA, while adhering almost perfectly to environmental constraints. We demonstrate the utility of our model through experiments on the nuScenes and PFSD datasets and provide an extensive benchmark analysis against the SOTA methods.

Keywords: Human Trajectory Prediction · Diffusion Model

1 Introduction

Recent Human Trajectory Prediction (HTP) [1,3,9,13,16,30] works have achieved great prediction accuracy by incorporating the stochastic nature of human trajectory; this means that HTP models sample multiple trajectory predictions and use the most accurate one for accuracy evaluation. HTP models also promote sample diversity to cover as many modes of movement as possible. However, SOTA models often sacrifice the feasibility and realism of the prediction in order to achieve the above-mentioned properties, failing to produce quality trajectory samples that consistently follow scene contexts. In this paper,

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/978-3-031-78110-0_25.

A. Antonacopoulos et al. (Eds.): ICPR 2024, LNCS 15329, pp. 382–397, 2025.
https://doi.org/10.1007/978-3-031-78110-0_25

we investigate how to achieve accurate and diverse predictions while making sure that the generated trajectories are also reasonable with respect to environmental constraints.

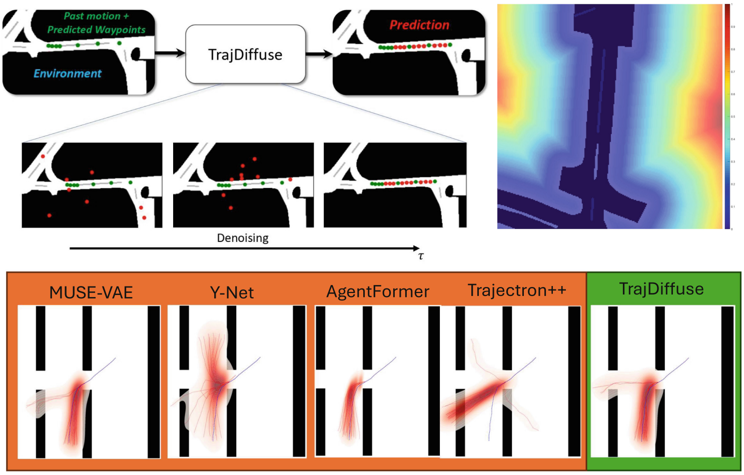


Fig. 1. Top Left: Illustration of the denoising trajectory prediction process. Green dots indicate the observed trajectory and the predicted way points. Red dots are the denoised prediction. **Top Right**: A map from nuScene dataset overlayed with the distance transform showing the distance to the navigable areas (roads) in the scene. The color scale represents the normalized distance from closest (blue) to the farthest (red). **Bottom**: Comparison against other SOTA methods on PFSD dataset (Color figure online)

To predict an accurate multimodal distribution of human trajectories, recent work has often adopted a planning-based approach by learning a probabilistic latent variable model of the agent's intent of movement [19,30]. Such formulation models the human decision-making process in which human agents often have a goal in mind and execute their movement based on their goals. These methods typically use a CVAE-based model to decode a latent variable to obtain the trajectory prediction. Despite the effectiveness of this strategy, we argue that these works often lack explicit control over the generation and decoding process.

Other non-probabilistic methods often rely on test-time sampling tricks or the anchoring approach [3,16]. These methods can achieve diverse sampling by using handcrafted criteria that promote the output trajectories to be distinct from each other. However, we have observed that the sampling trick will cause a trade-off between the diversity and the feasibility of predicted trajectories.

To address these issues, we propose a trajectory prediction method based on a guided conditional diffusion model. Compare with previous diffusion-based approaches [6,14,17], which use latent observation embedding as input, our method formulates the trajectory prediction problem as an interpolation between the agent trajectory history and the predicted agent motion intent, as illustrated on the top left of Fig. 1. By adopting the diffusion model framework, we gain the ability to directly and explicitly control the trajectory generation process via goal/waypoint conditioning and a guidance function that fuses the higher-level signal of an agent's intent with constraints such as the environmental awareness encoded on the top right of Fig. 1. Our method has shown better environmental understanding compared to other SOTA HTP methods as shown in both the bottom of Fig. 1 and experiments in the later sections.

Our contributions are summarized as follows: (1) We propose a novel planning-based trajectory prediction algorithm `TrajDiffuse` using a guided conditional diffusion model to obtain accurate and diverse trajectory predictions. (2) We propose an environment-based gradient guidance term that ensures that output trajectories are feasible and environment-compliant. (3) We demonstrate our design through experiments on two public datasets. We show that our method can produce an accurate trajectory prediction and achieve good compliance with environmental constraints. The source code can be found here: https://github.com/TL-QZ/TrajDiffuse.git

2 Related Work

Trajectory Prediction. Early works in trajectory prediction have focused on performing the trajectories prediction with sequence-to-sequence models such as the RNN-based Social-LSTM proposed by [1]. To capture the multimodality of human trajectories, recent work has often applied probabilistic generative frameworks for the prediction process. Social-GAN and Social-BiGAT [7,12] proposed to use the Generative Adverseral Network (GAN) [5] to generate multiple prediction outputs by repeatedly sampling inputs for the generative network. MUSE-VAE, Trajectron++ and Agentformer [13,22,30] applied the CVAE [24] for inference on the distribution of latent agent intents. For nonprobabilistic approaches, Y-Net [16] used a test-time sampling trick to achieve diverse trajectory predictions. MultiPath [3] leveraged a fixed set of state sequence anchors to generate diverse modes of trajectory predictions. GOHOME [4] uses lane information from HD-maps to generate heatmaps for intent prediction. Despite its effectiveness, such information are only viable for vehicle trajectory prediction, as pedestrian does not follows specific lanes. Besides, the evaluations of above-mentioned models often ignored feasibility and only promoted diversity and accuracy of the mode with least displacement from the ground truth, making generation unusable and unrealistic trajectory predictions. We have shown in our experiments that our model alleviates this problem by considering the environmental feasibility of each output prediction and guiding the sampling process with a map-based gradient correction term.

Diffusion Models. The recent success of diffusion models [8,23,27] in image and audio generation has shown their ability to generate high-quality, high-dimensional samples. The applications of diffusion models have also been extended to multimodal learning [18], sequence learning [20], and offline reinforcement learning tasks [10,29]. Recent works also attempt to apply diffusion for the HTP task. Gu et al. [6] proposed Motion Indeterminacy Diffusion (MID) that predicts the agent's trajectory by denoising diffusion of Gaussian noise to trajectory predictions using the observed trajectory embeddings as an additional condition to the model. This approach requires a large number of diffusion steps. Li et al. [14] and Mao et al. [17] proposed using a trajectory proposal rather than pure Gaussian noises as the start of the denoising process to reduce the number of steps required. Leapfrog Diffusion [17] generates the trajectory proposal from a learned deterministic trajectory initializer, while [14] used a CVAE-based module for this initialization. MotionDiffuser [11] focused on learning the joint distribution for motions of multiple agents to generate trajectory predictions without agent-agent collision. The above-mentioned diffusion-based HTP models all uses latent embedding of observed trajectories as condition and generate trajectories via denoising diffusion from either Gaussian noises or a full trajectory proposal. We argue that such a setting lacks explicit control over the generation process and does not offer a clear strategy that incorporates environmental information. Inspired by recent work [10] on the use of the diffusion model for goal-conditioned offline reinforcement learning, we propose a new planning-based HTP model using guided conditional denoising diffusion to solve human trajectory prediction as an impainting task by interpolating the agent's trajectory history and predicted agent motion intent. We also propose a map-based gradient guidance term to ensure that all generated trajectory samples are complying with the environmental constraint. We compare with the MID model and the results of our experiments demonstrate the effectiveness of our model in both predictive accuracy and quality performance. For other diffusion-based models, we were not able to included in the benchmark due to the source code being unavailable .

3 Proposed Method

We define the trajectory prediction problem as follows: Given the observed trajectory of an agent, denoted as $X = (x_1, \ldots, x_{T_o})$, where each $x_t \in \mathbb{R}^2$ represents the agent's 2D coordinates in timestep t within the observed frames T_o, as well as the semantic map M of the surrounding environment, our objective is to predict the future trajectory $Y = (y_{T_o+1}, \ldots, y_{T_o+T_p})$ for subsequent frames T_p. Here, $y_t \in \mathbb{R}^2$ denotes the 2D coordinates of the agent within the same coordinate system as X.

To achieve our objective, we formulate the trajectory prediction task as follows.

$$P(Y, G|X, M) = P(Y|G, X, M)P(G|X, M). \tag{1}$$

Here, the trajectory distribution is conditioned on an inferred agent intent G; we assume this intent to be the predicted long-term goal and short-term way-points the agent should follow, which makes $G \subset Y$. We used an off-the-shelf goal predictor to infer G. To model the conditional trajectory distribution $P(Y|G, X, M)$,

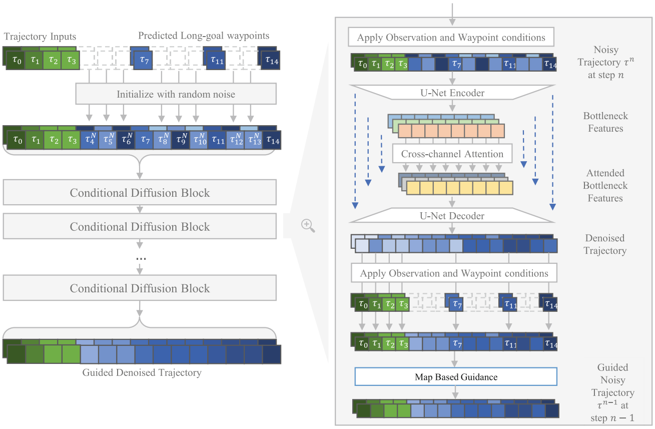

Trajectory Inputs
Predicted Long-goal waypoints
τ_0 τ_1 τ_2 τ_3 τ_7 τ_{11} τ_{14}
Initialize with random noise
τ_0 τ_1 τ_2 τ_3 τ_4^N τ_5^N τ_6^N τ_7 τ_8^N τ_9^N τ_{10}^N τ_{11} τ_{12}^N τ_{13}^N τ_{14}
Conditional Diffusion Block
Conditional Diffusion Block
...
Conditional Diffusion Block
Guided Denoised Trajectory
Apply Observation and Waypoint conditions
Noisy Trajectory τ^n at step n
U-Net Encoder
Bottleneck Features
Cross-channel Attention
Attended Bottleneck Features
U-Net Decoder
Denoised Trajectory
Apply Observation and Waypoint conditions
Map Based Guidance
Guided Noisy Trajectory τ^{n-1} at step $n-1$

Fig. 2. Details on `TrajDiffuse` Model Structure. Left: Prediction pipeline of the `TrajDiffuse` model. Here we represent the data as a two-channel one-dimensional signal; the two channels in the input and output correspond to the two dimensions of the position coordinate. Right: Illustration of the conditional denoising process inside a diffusion block. The input and output are conditioned on the observed trajectory and the predicted intents. The U-net-encoded bottleneck features are attended across the channels and decoded. The output is then conditioned by the observed trajectory history and the predicted waypoints.

we introduce a novel conditional diffusion-based trajectory prediction method called `TrajDiffuse`. This model takes advantage of the inferred intent of the agent, the observed trajectory, and a semantic scene map. It generates a trajectory that corresponds to the agent's intent, guided by a map-based guidance module that ensures that the generated trajectory adheres to environmental constraints.

In Sect. 3.1, we introduce the basic formulation of the diffusion model. Section 3.2 introduces the formulation of the model `TrajDiffuse` in terms of input/output representation, model formulation, model architecture, and training. Section 3.3 introduces a map-based guidance term which ensures that the sampled predictions adhere to environmental constraints. Finally, Sect. 3.4 introduces the sampling process during test time to generate trajectory predictions using the `TrajDiffuse` model.

3.1 Fundamentals of Denoising Diffusion Model

The Diffusion Model [8,23] models the data generation procedure as an iterative denoising process $P_\theta(\tau^{i-1}|\tau^i)$ for $i = N, \ldots, 0$ starting from $\tau^N \sim \mathcal{N}(0, I)$ sampled from a standard Gaussian distribution and τ^0 being the ground truth data instance. This process is the inverse of a forward procedure that gradually adds noise to the ground truth data instance based on a sequence of variance schedule hyperparameters $\alpha = \{\alpha_1, ..., \alpha_N\}$ which can be written as

$$q(\tau^t|\tau^{t-1}) \sim \mathcal{N}(\tau^t; \sqrt{\alpha_t}\tau^{t-1}, (1-\alpha_t)\mathbf{I}). \tag{2}$$

The reverse denoising model $P_\theta(\tau^{i-1}|\tau^i)$ is often modeled as

$$P_\theta(\tau^{i-1}|\tau^i) \sim \mathcal{N}(\mu_\theta(\tau^i, i), \sigma_q^2(i)\mathbf{I}). \tag{3}$$

Here, the mean function $\mu_\theta(\tau^i, i)$ is often parameterized with a neural network structure that takes the output of the previous denoising step and an embedding for the current denoising step index, and $\sigma_q^2(i)$ is a constant function of the scheduling hyperparameters α. We follow the definition in [15] for this diffusion process.

3.2 Conditional Diffusion Model for Trajectory Prediction

Trajectory Representation. We represent the input and output trajectories of the model $\tau \in \mathbb{R}^{(T_o+T_p)\times 2}$ as the concatenation of the observed trajectory X and the predicted trajectory $\hat{Y}$. Then, the trajectory output from denoising step i will have the form

$$\tau^i = (x_1, \dots, x_{T_o}, \hat{y}^i_{T_o+1}, \dots, \hat{y}^i_{T_o+T_p-1}, \hat{y}_{T_o+T_p})'. \tag{4}$$

There are two notions of timesteps, the index of denoising steps and the number of trajectory frames. We will use superscript to represent the former and subscript to represent the latter. This means that $\hat{y}^i_{T_o+1}$ represents the agent's 2D coordinates in the $T_o + 1$ frame from the output of the ith denoising step.

We assume the predicted intent G consists of a predicted goal $\hat{y}_{T_o+T_p}$ and S waypoints $\hat{y}_{w_1} \dots \hat{y}_{w_S}$, where $w_s \in \{T_o+1, \dots, T_o+T_p-1\}$ for all $s \in \{1, \dots, S\}$, of the form

$$G = \{\hat{y}_{w_1} \dots \hat{y}_{w_S}, \hat{y}_{T_o+T_p}\}. \tag{5}$$

Model Formulation. Inspired by the image-inpainting task presented in [23], our `TrajDiffuse` model $P(Y|G, X, M)$ considers the trajectory prediction task as an interpolation for the observed trajectory history and the predicted waypoints and the end goal. We use the diffusion model to perform iterative denoising on the coordinates between the observed history and the predicted goal and waypoints. Our model has the form

$$P(Y|G, X, M) = P(\tau^0) = \int P(\tau^N) \prod_{i=0}^{N-1} P(\tau^i|\tau^{i+1}) d\tau^{1:N}. \tag{6}$$

For each denoising step i, we fix the elements in τ^i corresponding to the observed trajectory with the observed coordinate sequence $(x_1, \dots, x_{T_o})$ and the elements corresponding to the predicted goal and the waypoints in the predicted intent G to obtain a noisy conditioned trajectory $\tau^{i'}$. We then feed the $\tau^{i'}$ into the next denoising step $i-1$ and sample the next denoised trajectory τ^{i-1} based on the distribution in (3). Figure 2 illustrates how the input to each diffusion step is conditioned by the observed trajectory and the predicted agent intent.

Architecture. Following [15], we further parameterize the mean function $\mu_\theta(\tau^i, i)$ as

$$\mu_\theta(\tau^i, i) = \mu(\tau_\theta(\tau^i, i), \alpha), \tag{7}$$

where $\tau_\theta(\tau^i)$ is a neural network that predicts the ground truth trajectory with the noisy trajectory as its input. We use a U-Net-based [21] structure along with an attention module [28] for $\tau_\theta(\tau^i)$. We considered the two-dimensional trajectory in the form of a two-dimensional coordinate sequence as a 2-channel 1-dimensional image with only one temporal dimension. Therefore, we use groups of 1-dimensional convolutional neural network blocks along with residual connections for the encoding and decoding modules for the U-Net structure, following the model architecture design in [10]. The encoded U-Net feature, which has a dimension of $C \times W$, does not contain any information across the channels during the encoding process. Therefore, we feed the encoded feature into a cross-channel attention layer to resolve this issue, as suggested in [10]. The cross-attended feature is the input into the decoding blocks to generate the denoised trajectory. Figure 2 illustrates the structure introduced here.

Training. Like standard denoising diffusion models, to train `TrajDiffuse`, we minimize the KL divergence between our denoising model $P_\theta(\tau^{i-1}|\tau^i)$ and the ground-truth denoising distribution $q(\tau^{i-1}|\tau^t, \tau^0)$ for each denoising step $i \in [1, N]$. This KL divergence becomes the L_2 norm between the predicted mean function $\mu_\theta(\tau^i, i)$ and the ground-truth mean function $\tilde{\mu}(\tau^t, \tau^0)$. Based on the repprameterization in (7), we can further simplify this and form our loss function as

$$\mathcal{L} = \mathbb{E}_{i \sim U\{1,N\}} \left[\mathbb{E}_{q(\tau^i|\tau^{i-1})} \left[\frac{\lambda(\alpha)}{2\sigma_q^2(t)} \|\tau^\theta(\tau^i, t) - \tau^0\|_2^2 \right] \right]. \tag{8}$$

The λ (α) is a function of the predefined noise schedule hyperparameter α. The complete derivation follows the review [15] and the original DDPM paper [8]. We also provide a brief derivation in the Supplementary Material.

3.3 Map-Based Gradient Guidance Module

To achieve the goal of generating trajectory predictions that strictly adhere to environmental constraints, we introduce a novel map-based gradient guidance term for the denoising trajectory prediction process, illustrated as the Map Based Guidance module in Fig. 2. Given the semantic map M, we first extract the binary navigability map M_b and perform the distance transform $D(M_b)$ to obtain the distance map M_d. Each element $M_d(i, j)$ represents the distance to the closest navigable pixel, as demonstrated on right of Fig. 1. We then calculate the image gradient ∇M_d with respect to each pixel coordinate.

Directly using the gradient to guide the denoised trajectory may cause the coordinates of different frames to drift toward opposing directions, causing impossible predictions. Therefore, we propose an algorithm to guide the trajectory iteratively through the trajectory sequence. For each frame in the trajectory, we will perform a gradient descent using the image gradient ∇M_d to ensure that

Algorithm 1: Map-based Guidance $\mathcal{H}(\nabla M_d, \tau^i)$

```
Input: denoised trajectory $\tau^i$
Output: map-based guidance term
$\tau^{i*} \leftarrow \tau^i$;
for $f = T_o + 1$ to $T_o + T_p$ do
    for $k = K, \ldots, 1$ do
        // K times of this gradient descent step
        $\tau^{i*}[f:] \leftarrow \tau^{i*}[f:] + \nabla M_d(\tau^{i*}[f])$;
    end
end
return $\tau^{i*} - \tau^i$
```

the coordinates of the current frame end up in a navigable area; we then also update all the positions of the later frames relative to the updated location of the current frame, as illustrated in Algorithm 1. We continue to perform this update iteratively for all time steps $\{t | t = T_o + 1, \ldots, T_o + T_p\}$.

3.4 Test Time Sampling

During test time, we assume that an agent has an observed history X, the semantic scene map M, and K sets of predicted agent intent $G = \{G_1, \ldots, G_K\}$. For each set G_k, we sample a trajectory prediction through the conditional denoising diffusion process guided by the map-based guidance $\mathcal{H}$ defined in Sect. 3.3. The pseudocode for the guided prediction pipeline is given in the supplementary.

4 Experiments

Section 4.1 introduces the datasets, evaluation metrics, and benchmark settings. Section 4.2 quantitatively compares the SOTA models and the `TrajDiffuse` model. Section 4.3 performs qualitative analysis for the SOTA models and our `TrajDiffuse` model. In Sect. 4.4, we present an ablation study for the map-based guidance module. We also performed an analysis of the inference speed compared to other SOTA models, included in the supplementary section G.

4.1 Preliminaries

Datasets. We use two publicly available datasets for our benchmark experiment. The **nuScenes** dataset, first introduced in [2], is a large-scale vehicle trajectory dataset, consisting of 1000 driving scenes and provides HD semantic maps. There are multiple benchmark configurations for this dataset and we follow [13,30] and used their training and testing splits for the nuScenes prediction challenge. The **PFSD** dataset introduced in [13] features simulated trajectories within a group of synthetic path-finding environment layouts proposed in

[26]. The non-navigable areas are designed to be more complex for navigation. Despite having different agent types (vehicle and pedestrian, respectively), both benchmark offers strict navigability constraint, making them great benchmark platforms testing the models' ability generating realistic predictions.

Metrics. For the benchmark evaluation, we compute the minimum average displacement error $\mathbf{ADE_k}$ and the final displacement error $\mathbf{FDE_k}$ of the K trajectory samples for each agent compared with the ground truth trajectory. We also adopt the Kernel Density Estimated-based Negative Log Likelihood (**KDE NLL**) proposed in [9], which evaluates the fit of the model. We evaluate the environment understanding of the prediction models using the Environmental Collision-Free Likelihood (**ECFL**) proposed in [25], which measures the probability that a predicted trajectory free of environmental collision. Finally, we quantify the diversity of the prediction outputs using the Multiverse Entropy **(MVE)** proposed in [25] quantifying the diversity of the trajectory predictions. We provides a detailed definition of each metric in the supplementary sections of the paper.

Implementation Details. We use the CVAE-based Macro-stage of the MUSE-VAE model [13] to predict intent of an agent. We also point out that, for the following experiments, `TrajDiffuse` and MUSE-VAE share the same sets of predicted waypoint sets. Other implementation details are presented in the supplementary.

4.2 Quantitative Analysis

For the quantitative experiment, we perform the trajectory prediction task on the two datasets mentioned above. We compared the performance of `TrajDiffuse` against the Trajectron++ (T++) [22], AgentFormer (AF) [30], Y-net [16], MUSE-VAE (MUSE) [13] and motion indeterminacy diffusion (MID) [6]. Both PFSD and nuScenes provide rasterized global scene maps, and we provide a local view of the maps for all the methods benchmarked following the experiment in [13]. The original MID model does not utilize map information, so we also compare with a modified MID that incorporates a map embedding for fair comparison. Since MID uses T++'s encoder for observed trajectories and social interactions, we use T++'s map encoder for the modified MID. We point out this is different from our map guidance module, but an adaption based on the design choice of MID model itself. We trained all models from scratch except for MUSE and Agentformer, where we used their provided pre-trained weight on the nuScenes dataset and we used MUSE-VAE's pretrained weight on PFSD dataset. Some benchmark models consider multi-agent settings, using context from other agents to condition independent or joint predictions of all agents' trajectories in the scene. We argue that such a difference does not significantly affect our conclusions, and we provide a more detailed discussion in the supplementary material.

Table 1. Quantative Results on PFSD and nuScenes datasets

(a) Results on PFSD with $K = 20$ with $t_{obs} = 3.2$s (8 frames) and $t_{pred} = 4.8$s (12 frames). Errors are in meters. The best performance is boldfaced and the 2nd place is marked as blue. Numbers in parenthesis indicate the ranking for the score.

Model	ADE ↓	FDE ↓	NLL ↓	ECFL ↑	MVE ↑
T++	0.20 (7)	0.42 (7)	2.24 (7)	85.00 (7)	**1.13** (1)
AF	0.11 (6)	0.17 (5)	**1.93** (1)	93.76 (4)	0.67 (7)
Y-net	0.07 (3)	0.12 (3)	1.98 (3)	94.16 (3)	0.79 (6)
MUSE	**0.05** (1)	**0.09** (1)	1.95 (2)	97.08 (2)	0.92 (4)
MID	0.09 (4)	0.16 (4)	2.00 (5)	88.72 (6)	0.93 (3)
MID w/Map	0.10 (5)	0.19 (6)	2.00 (5)	90.41 (5)	0.86 (5)
`TrajDiffuse`	0.06 (2)	**0.09** (1)	1.98 (3)	**99.62** (1)	1.08 (2)

(b) Result on nuScenes with $K = 5$ and $K = 10$ with $t_{obs} = 2$s (4 frames) and $t_{pred} = 6$s (12 frames). Errors are in meters. The best performance is boldfaced and the 2nd place is marked as blue. Numbers in parenthesis indicate the ranking for the score.

K	Model	ADE ↓	FDE ↓	NLL ↓	ECFL ↑	MVE ↑
5	T++	2.51 (7)	5.57 (6)	11.66 (7)	81.66 (4)	0.46 (6)
	AF	1.86 (4)	3.89 (4)	6.94 (3)	84.66 (3)	0.38 (7)
	Y-net	1.63 (2)	2.86 (3)	7.13 (4)	76.61 (5)	0.68 (3)
	MUSE	**1.37** (1)	2.84 (2)	**5.76** (1)	89.30 (2)	0.65 (4)
	MID	2.38 (5)	5.54 (5)	9.33 (5)	69.23 (6)	**0.81** (1)
	MID w/Map	2.42 (6)	5.61 (6)	9.51 (6)	68.72 (7)	**0.81** (1)
	`TrajDiffuse`	1.67 (3)	**2.73** (1)	6.85 (2)	**99.15** (1)	0.61 (5)
10	T++	1.92 (5)	4.01 (5)	8.20 (7)	81.25 (4)	0.57 (6)
	AF	1.45 (4)	2.86 (4)	5.67 (4)	84.26 (3)	0.42 (7)
	Y-net	1.32 (1)	2.05 (2)	5.60 (3)	70.71 (5)	**1.03** (3)
	MUSE	**1.10** (1)	2.11 (3)	**4.61** (1)	89.26 (2)	0.79 (4)
	MID	1.93 (6)	4.29 (7)	7.42 (6)	68.97 (6)	1.00 (2)
	MID w/Map	1.96 (7)	4.28 (6)	7.41 (5)	68.40 (7)	1.00 (2)
	`TrajDiffuse`	1.41 (3)	**2.02** (1)	5.33 (2)	**99.08** (1)	0.74 (5)

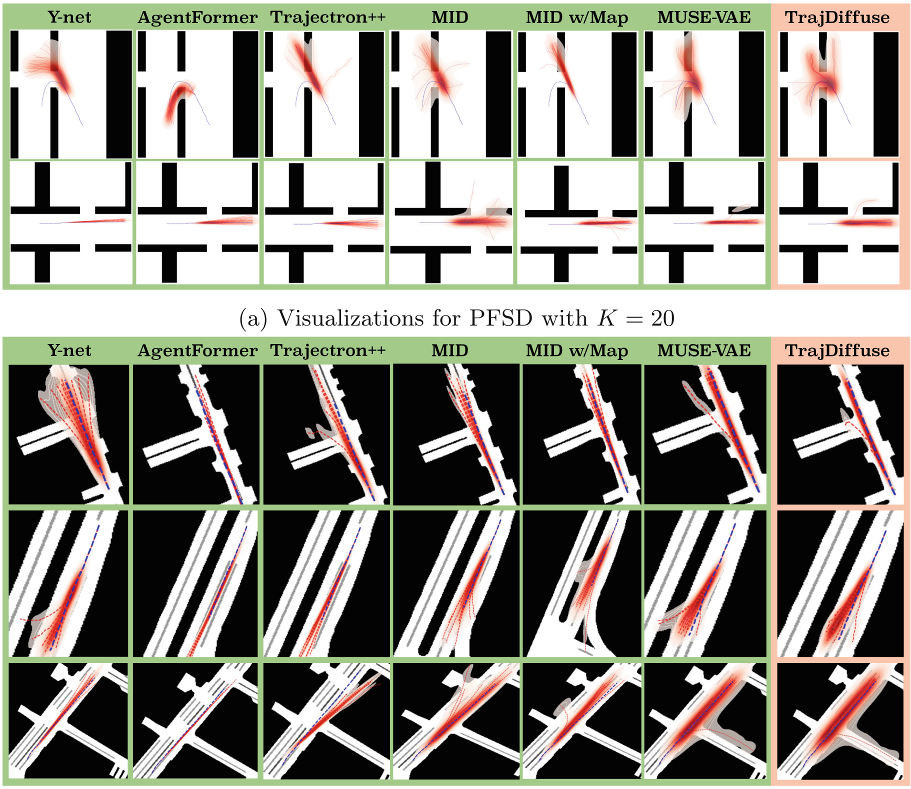
PFSD Dataset. Table 1a summarizes the results in the PFSD dataset. For the PFSD dataset, we use 3.2 s (8 frames) observations and predict 4.8 s (12 frames) into the future. We chose to sample $K = 20$ samples for the PFSD dataset to consider the inherent multimodal nature of the human trajectory. For the ADE_{20} score, `TrajDiffuse` was able to achieve the second best among all the benchmarked models. For the FDE_{20} score, `TrajDiffuse` achieves the overall best along with the MUSE-VAE model. The two displacement errors measure the prediction accuracy of the best sample and we have shown that `TrajDiffuse` can achieve SOTA performance in terms of these two metrics. For the KDE NLL metric, `TrajDiffuse` also matches the SOTA methods, meaning that the K samples generated could reflect the distribution of the ground truth data distribution. `TrajDiffuse` model is capable of surpassing all SOTA models in the ECFL experiment. The `TrajDiffuse` model can achieve an almost perfect prediction in terms of complying with the environmental constraint. This indicates that given sufficient predicted waypoints, `TrajDiffuse` is capable of generating predictions that are accurate and realistic. For the MVE, `TrajDiffuse` is ranked second. This indicates that `TrajDiffuse` is capable of generating diverse trajectory predictions. Compared to MID, another diffusion-based model, `TrajDiffuse` achieved better performance in all metrics. For the PFSD data set, after adding map embedding, MID achieves a lower MVE score, causing the predictions to be less diverse. This is also demonstrated in the qualitative analysis.

nuScenes Dataset. For the nuScenes dataset, we follow the configurations of previous works, observing 2 s (4 frames) of past trajectories and predicting 6 s (12 frames) into the future. We experiment with two K settings for this dataset. Table 1b shows the result. For the $K = 5$ case, our model was able to achieve the best FDE_5 score and matches the SOTA ADE_5 performance. This indicates that the model can take advantage of the accurate long-term prediction and generate a reasonable and accurate trajectory. For the KDE NLL, `TrajDiffuse` is able to achieve the second best. For the ECFL score, `TrajDiffuse` can achieve the best among all models and achieve an increase of almost 10%. For case $K = 10$, all models achieve better displacement error performance, with MUSE having the best ADE_{10} score and `TrajDiffuse` having the best FDE_{10} score. `TrajDiffuse` was also able to achieve SOTA performance on the KDE NLL metric. For the ECFL, `TrajDiffuse` again achieves the best among the benchmark models and achieves almost perfect predictions it terms of following the environmental constraint. For the MVE metric, `TrajDiffuse` is ranked 5th overall. However, compared to the top three ranking MID, Y-net and MUSE-VAE, `TrajDiffuse` is capable of generating more feasible predictions while still having comparable MVE values. On the other hand, the top-ranking Y-net and MID sacrifice the feasibility in terms of ECFL to achieve more diversity, causing most of their prediction to be unrealistic. Although MUSE-VAE ranked the highest on average, its generated trajectories often violated environmental constraints, making unrealistic predictions, as can be seen in the qualitative analysis in the next section. The MID model struggled to perform on the vehicle-focused nuScenes dataset. Despite ranking first for the MVE metric for diversity, the model fails to produce accurate and feasible trajectory predictions. The experiments show that diffusing from raw Gaussian noise and using only trajectory history embedding as conditions is not as effective compared with our setting of interpolation via diffusion.

4.3 Qualitative Analysis

We provide a qualitative analysis with visualizations of the prediction output of each model to better illustrate the behavior and characteristics of these models.

Figure 3a shows two instances of the PFSD experiment. For the PFSD dataset, the `TrajDiffuse` model is capable of generating diverse paths that are also able to adhere to the environmental constraint. Here, we point out again that we use the same intent prediction as the MUSE-VAE output. Since MUSE-VAE does not directly use the predicted goal and waypoints, it will still generate trajectories that cause a collision with the environment. Our guided denoising inpainting approach is able to address this issue while still maintaining the prediction accuracy. Y-net test time sampling trick helps the model generate a diverse set of predictions; however, this sampling process is not data-driven and generates trajectories that ignore the environmental constraint. The AgentFormer model is able to generate trajectories that cause no collision with the environment in one of the instances, however, the outputs lack diversity. For the other instance, AgentFormer is able to predict the correct intent of the agent; however, the trajectories all violate the environmetnal constaint and the predictions also lack diversity. Trajectron++ model suffers a lack of diversity issue in one of the cases shown here, and in the other instance the model also generates many trajectories which ignore the environmental constraint. The MID model (in both settings) behave similarly to the Trajectron++, as they share the same encoder structure. The model is able to achieve better environmental understanding with the addition of map embedding; however, it still produces



(a) Visualizations for PFSD with $K = 20$

(b) Visualizations for nuScenes with $K = 10$

Fig. 3. Visualizations for qualitative analysis on PFSD and nuScenes datasets. Each column contains visualizations of an agent's trajectories predicted by the model indicated at the top of the column. Each row corresponds to the agent with identical initial conditions and identical prior motion history. Blue dashed lines denote the ground-truth trajectories. Red dashed lines are predicted trajectories. (Color figure online)

trajectories violating the map constraints, and the addition of map embedding also causes the model to produce less diverse predictions in both instances.

Figure 3b presents three instances from the nuScenes dataset. Due to the dataset's focus on vehicle trajectory in traffic scenarios, the environmental constraint is stricter, offering limited navigable areas. Here, the `TrajDiffuse` model is capable of producing accurate and diverse trajectories and staying within navigable areas. `TrajDiffuse` and MUSE-VAE also share the predicted waypoints in this experiment. However, we see here that MUSE will generate off-road trajectories, which are likely due to inaccurate intent predictions. Meanwhile, the `TrajDiffuse` model with its map-based guidance is able to correct those trajectories and makes reasonable predictions. For Y-net outputs, the disadvantage of the test-time sampling trick is shown in this kind of narrow environment.

Y-net outputs have shown a trade-off between diversity and environmental compliance, where many outputs go off-road. The sequence-to-sequence setting of AgentFormer and Trajectron++ was able to generate more output that stays on-road. However, these predictions often fail to reach the ground truth goal and overshoot beyond the truth trajectory, causing the prediction accuracy problem. MID also suffers from this issue of inaccurate predictions, and the model also often goes off road given the narrower navigable areas. With the help of map embedding, the model does have lower environmental violation; however, it still suffers in accuracy, and the model also produces less diverse outputs for these three instances.

From the qualitative analysis, we have shown that the `TrajDiffuse` offers a better solution for using the predicted intent than the MUSE-VAE model and renders more realistic and diverse predictions than MID Y-Net, AgentFormer and Trajectron++ models.

4.4 Ablation Study

Table 2a presents an ablation study of the map-based guidance sampling on the nuScenes dataset with $K = 10$ samples. We note that the map-based guidance not only improves the ECFL of the model output; it also improves the accuracy metrics. The MVE value decreases slightly; however the MVE metric should be considered along with other metrics and this indicates the guidance term helps the model to generate higher quality trajectory predictions while maintaining the diversity. We also note that using the heatmap based MUSE Macro-stage for intent prediction indeed helps the `TrajDiffuse` achieves a good starting base line for ECFL, however, we see that using the same sets of predicted waypoints, `TrajDiffuse` without map guidance still achieves better ECFL score than MUSE and both `TrajDiffuse` settings also achieve better FDE score compare with MUSE. This is due to the non-autoregressive setting of our diffusion based interpolation, which directly uses the environmentally-compliant waypoints as the final output. On the other hand, MUSE-VAE's micro-stage only uses these predicted waypoints as reference for its auto-regressive RNN based prediction module and generate brand new trajectory predictions that often fails to maximizes the advantage of the environment-aware Macro-stage predictions.

We provide another ablation study in Table 2b on the number waypoints used for prediction on the nuScenes dataset. With only one waypoint (final goal), the model yields slightly decreased performance. However, for other choices, the model performance increases and does not drastically vary. Hence, the model needs both the intermediate waypoints and the goal point, but it is robust against a different number of waypoints. We chose 3 waypoints following MUSE-VAE.

5 Discussion and Conclusion

In this paper, we introduced `TrajDiffuse`, a guided conditional diffusion model for trajectory prediction. By framing the prediction task as denoising interpolation of the observed trajectory and predicted waypoints, our model is able to

Table 2. Two abalation studies

(a) Ablation study of the map-based guidance on nuScenes dataset with $K = 10$.

Model	ADE ↓	FDE ↓	KDE NLL ↓	ECFL ↑	MVE ↑
MUSE-VAE	1.10	2.11	4.61	89.26	0.79
`TrajDiffuse` w/o Guidance	1.42	2.04	5.40	92.07	0.77
`TrajDiffuse` w Guidance	1.42	2.02	5.33	99.08	0.75

(b) Ablation study for different number of waypoints on nuScenes with $K = 5$ with $t_{obs} = 2$s (4 frames) and $t_{pred} = 6$s (12 frames). Errors are in meters; map-guide vs. w/o map-guide

#Waypoints	ADE ↓	FDE ↓	NLL ↓	ECFL ↑	MVE ↑
1	1.74/1.77	2.75/2.78	7.18/7.30	99.16/89.12	0.57/0.60
2	1.68/1.69	2.76/2.80	7.12/7.22	99.15/91.64	0.60/0.62
3	1.67/1.68	2.73/2.76	6.85/6.94	99.15/92.10	0.61/0.64
4	1.66/1.67	2.77/2.79	7.15/7.20	99.02/92.24	0.60/0.62

achieve a SOTA performance in accuracy and diversity measure while surpassing the existing method's ability to follow the critical environmental constraints.

Several other future directions can be considered. Our `TrajDiffuse` framework makes it possible to integrate other dynamic scene elements and HD-maps that have become more prevalent in recent years. Alternate approaches that incorporate the intent of other agents in the scene to model agent-agent interactions may similarly and readily benefit from the `TrajDiffuse` framework.

Our method leverages the predicted agent intent in the form of trajectory waypoints; in this work, we employ a CVAE-based backbone of the MUSE-VAE model [13] for that task. Although we demonstrated that the feasibility of the `TrajDiffuse` prediction exceeds that of MUSE-VAE, other waypoint prediction models could be used to create successful pipeline designs, opening an interesting direction for further investigation.

Acknowledgement. This project is supported in part by National Science Foundation Grant Awards #1955404 and #1955365.

References

1. Alahi, A., Goel, K., Ramanathan, V., Robicquet, A., Fei-Fei, L., Savarese, S.: Social lstm: human trajectory prediction in crowded spaces. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 961–971 (2016)
2. Caesar, H., et al.: nuscenes: a multimodal dataset for autonomous driving. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 11621–11631 (2020)
3. Chai, Y., Sapp, B., Bansal, M., Anguelov, D.: Multipath: Multiple probabilistic anchor trajectory hypotheses for behavior prediction. arXiv preprint arXiv:1910.05449 (2019)
4. Gilles, T., Sabatini, S., Tsishkou, D., Stanciulescu, B., Moutarde, F.: Gohome: graph-oriented heatmap output for future motion estimation. In: 2022 International Conference on Robotics and Automation (ICRA), pp. 9107–9114. IEEE Press (2022)
5. Goodfellow, I., et al.: Generative adversarial nets. Adv. Neural Inform. Process. Syst. **27** (2014)
6. Gu, T., et al.: Stochastic trajectory prediction via motion indeterminacy diffusion. In: CVPR (June 2022)
7. Gupta, A., Johnson, J., Fei-Fei, L., Savarese, S., Alahi, A.: Social gan: socially acceptable trajectories with generative adversarial networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2255–2264 (2018)

8. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. Adv. Neural. Inf. Process. Syst. **33**, 6840–6851 (2020)
9. Ivanovic, B., Pavone, M.: The trajectron: probabilistic multi-agent trajectory modeling with dynamic spatiotemporal graphs. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 2375–2384 (2019)
10. Janner, M., Du, Y., Tenenbaum, J., Levine, S.: Planning with diffusion for flexible behavior synthesis. In: International Conference on Machine Learning (2022)
11. Jiang, C."., et al.: Motiondiffuser: controllable multi-agent motion prediction using diffusion. In: CVPR (2023)
12. Kosaraju, V., et al.: Social-bigat: multimodal trajectory forecasting using bicycle-gan and graph attention networks. Adv. Neural Inform. Process. Syst. **32** (2019)
13. Lee, M., Sohn, S.S., Moon, S., Yoon, S., Kapadia, M., Pavlovic, V.: Muse-vae: multi-scale vae for environment-aware long term trajectory prediction. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2221–2230 (2022)
14. Li, Z., et al.: A multi-modal vehicle trajectory prediction framework via conditional diffusion model: a coarse-to-fine approach. Knowl.-Based Syst. **280**, 110990 (2023)
15. Luo, C.: Understanding diffusion models: A unified perspective. arXiv preprint arXiv:2208.11970 (2022)
16. Mangalam, K., An, Y., Girase, H., Malik, J.: From goals, waypoints & paths to long term human trajectory forecasting. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 15233–15242 (2021)
17. Mao, W., et al.: Leapfrog diffusion model for stochastic trajectory prediction. In: CVPR (2023)
18. Nichol, A., et al.: Glide: towards photorealistic image generation and editing with text-guided diffusion models. arXiv preprint arXiv:2112.10741 (2021)
19. Pang, B., Zhao, T., Xie, X., Wu, Y.N.: Trajectory prediction with latent belief energy-based model. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 11814–11824 (2021)
20. Rasul, K., Seward, C., Schuster, I., Vollgraf, R.: Autoregressive denoising diffusion models for multivariate probabilistic time series forecasting. In: International Conference on Machine Learning, pp. 8857–8868. PMLR (2021)
21. Ronneberger, O., Fischer, P., Brox, T.: U-Net: convolutional networks for biomedical image segmentation. In: Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F. (eds.) MICCAI 2015. LNCS, vol. 9351, pp. 234–241. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-24574-4_28
22. Salzmann, T., Ivanovic, B., Chakravarty, P., Pavone, M.: Trajectron++: dynamically-feasible trajectory forecasting with heterogeneous data. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J.-M. (eds.) ECCV 2020. LNCS, vol. 12363, pp. 683–700. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-58523-5_40
23. Sohl-Dickstein, J., et al.: Deep unsupervised learning using nonequilibrium thermodynamics. In: ICML, pp. 2256–2265. PMLR (2015)
24. Sohn, K., Lee, H., Yan, X.: Learning structured output representation using deep conditional generative models. Adv. Neural Inform. Process. Syst. **28** (2015)
25. Sohn, S.S., et al.: A2x: an agent and environment interaction benchmark for multimodal human trajectory prediction. In: Proceedings of the 14th ACM SIGGRAPH Conference on Motion, Interaction and Games, pp. 1–9 (2021)
26. Sohn, S.S., Zhou, H., Moon, S., Yoon, S., Pavlovic, V., Kapadia, M.: Laying the foundations of deep long-term crowd flow prediction. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J.-M. (eds.) ECCV 2020. LNCS, vol. 12374, pp. 711–728. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-58526-6_42

27. Song, Y., Ermon, S.: Generative modeling by estimating gradients of the data distribution. Adv. Neural Inform. Process. Syst. **32** (2019)
28. Vaswani, A., et al.: Attention is all you need. Adv. Neural Inform. Process. Syst. **30** (2017)
29. Wang, Z., Hunt, J.J., Zhou, M.: Diffusion policies as an expressive policy class for offline reinforcement learning. arXiv preprint arXiv:2208.06193 (2022)
30. Yuan, Y., Weng, X., Ou, Y., Kitani, K.M.: Agentformer: agent-aware transformers for socio-temporal multi-agent forecasting. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 9813–9823 (2021)

Principal Graph Neighborhood Aggregation for Underwater Moving Object Detection

Meghna Kapoor, Badri Narayan Subudhi(✉), Vinit Jakhetiya, and Ankur Bansal

Indian Institute of Technology Jammu, Jammu and Kashmir, India
{meghna,vinit.jakhetiya,ankur.bansal}@iitjammu.ac.in,
subudhi.badri@gmail.com

Abstract. Identifying and tracking mobile objects in an underwater environment is a challenging task. Traditional methods cannot differentiate the object from the background due to the losses induced by inherent properties of light. In this regard, researchers across the globe developed several deep learning models that adhered to convolutional kernels and their modified forms to tackle the same. However, such kernels incur losses due to uncertain, fuzzy, and poorly defined boundaries of objects inside the water. Recent advancements in graph learning have eliminated the loss typically associated with convolutional kernels, creating new opportunities for moving object identification. This study presents an end-to-end moving object detection architecture to analyze intricate underwater scenes. We adhered to a ResNet-50 backbone in the proposed architecture to project the video frame to feature space. Graph learning is used to retain the structural information of the object by projecting from feature space to graph space. Multiple aggregators facilitate the seamless transfer of information among neighbouring nodes, alleviating noise induced by deep architectures. The refactored latent vector is transformed to image space to detect the moving object(s) from the given scene. The proposed method is evaluated against twenty-four state-of-the-art algorithms on the benchmark datasets, outperforming all existing methods.

Keywords: Underwater object detection · Graph learning · Deep learning · Graph aggregation

1 Introduction

The advancement in underwater navigation has led to a greater need for detecting moving objects to facilitate automated navigation [22]. Despite significant advancements in terrestrial object recognition, detecting underwater objects remains an area for improvement. Conventional terrestrial approaches [23,41] could be more effective in underwater scenarios due to the challenges presented by the dynamics of the underwater environment. The objects are camouflaged and occluded, and their intricacies are obscured, making it challenging to discern the moving object. In addition, when light traverses in the underwater scenario, it loses some energy due to the intrinsic property of light. Hence, the underwater images are degraded in terms of colour and texture information,

A. Antonacopoulos et al. (Eds.): ICPR 2024, LNCS 15329, pp. 398–412, 2025.
https://doi.org/10.1007/978-3-031-78110-0_26

making it difficult to distinguish between foreground and background [31]. Further, the fish manoeuvre through the water with swift and sudden movements, making monitoring challenging. The challenges posed by underwater are more severe than those in terrestrial environments. Therefore, detecting moving objects in underwater environments is a relatively uncommon topic in the literature.

Recent studies [29,30,37,38] indicate notable progress in underwater surveillance by utilizing various features of objects. However, these methods fail due to degradations offered by the underwater environment. The loss of colour and texture information in underwater environments makes it difficult to rely on these features. Further, authors [1,25] have utilized deep-learning techniques to detect the foreground from the background, taking advantage of the progress in deep learning. Nevertheless, these methods are ineffective in preserving minute details of the object. One possible explanation is that noise is amplified by applying a spatial kernel in the non-Euclidean domain during the convolution operation and by the noise introduced by the max pooling operation.

The graph networks are more reliable in retaining information irrespective of the topology of the data. Hence, it has become popular in various domains like action recognition [33], cancer detection [43], etc. However, the spatial relationship extracted by the convolutional network is paramount in image processing tasks like object detection. Moreover, CNNs are less computationally expensive than graph networks. Hence, combining both networks can be utilized to preserve features better. Kapoor et al. [16] have introduced a graph-sage-based approach to detect moving objects in a scene using graph-refactoring in the latent space. This approach uses CNN for feature extraction and GraphSage to refactor the information shared by latent vector elements. Further, the author tests various aggregators in the sampled neighbourhood. The transfer of information among the nodes depends upon the number of nodes in the neighbourhood and the aggregator utilized to combine the information among neighbours. However, Xu et al. [40] demonstrated that the mean and max aggregators cannot differentiate between features with the same characteristics but different cardinalities. The graph structure information cannot be captured by mean, and max operators used by [16] to accurately refactor the latent vector. Corso et al. [3] proposed a principal neighbourhood aggregation with multiple aggregators and scalers to combine the information among neighbours, which can be utilized to refactor the feature vector in high-dimensional space.

This study presents an underwater moving object detection module with an end-to-end encoder-decoder architecture. The feature extraction module preserves the spatial correlation between the nodes and extracts the features in a high-dimensional latent space that differentiates the object from the background. The spatial correlations in the latent space are elusive. Therefore, we introduce a graph refactoring module to remove unwanted noise and retain the global contextual information. The graph refactoring module, positioned between the encoder and decoder architecture, preserves the intricate details while removing any distortion induced by the convolution kernel and max pooling. The latent vector is mapped onto the graph space, facilitating the exchange of information among the graph nodes. The performance of the graph is contingent upon the exchange of information across nodes. Therefore, using the principal aggregation approach, we utilize several aggregators and scalars to merge the information across the nodes. Aggregators and scalers collect and organize information from neighbouring nodes and pass it to the following graph layer. Following refactoring, the graph is transformed to generate a refactored latent vector. The decoder network converts the latent vector into the image domain and detects the moving object(s) present in the scene.

The main contributions of the article are as follows:

- A principal-neighborhood aggregator technique is used to restructure the latent space node relationship.
- A diverse range of aggregator functions is utilized with distinct scalars to preserve relevant information and eliminate noise.
- The proposed method was evaluated against twenty-four state-of-the-art techniques, and our method outperformed the existing methods in terms of performance.

The rest of the paper is organized as follows. The related works are discussed in Sect. 2. Section 3 demonstrates the proposed work and basic concepts. The results and discussion are presented in Sect. 4. Finally, Sect. 5 contains the conclusion of the proposed method.

2 Related Works

The state-of-the-art methods can be classified into two categories: traditional methods and learning-based methods. The traditional methods use image features and descriptors to identify the object(s) in the given scene. In contrast, deep learning-based methods don't require any pre-requisite prior knowledge or handcrafted features.

2.1 Traditional Methods

Traditionally, a moving object can be identified by analyzing the neighbouring information to detect any changes in the region across consecutive frames. A simple approach is subtracting the subsequent frames to detect local change across the frames [26,29]. Nevertheless, these strategies are ineffective when faced with sudden changes in motion and appearance of objects. Additionally, dynamic background conditions caused by algae and seaweed present a challenge, which complicates moving object detection. A simple way to incorporate the dynamic background is to model the background with a Gaussian Mixture Model (GMM) [15,27,28,30,36,45]. Further, the object can be identified by utilizing its attributes in the image, such as colour and texture information, as the object's appearance differs from that of the background. Authors in [37,38] suggested techniques that rely on identifying the object(s) through the object's features, like colour and texture characteristics. Nevertheless, the presence of colour is not reliable in underwater environments due to image deterioration. Further, Palazzo et al. [24] proposed a foreground and background modelling methodology. The method utilizes kernel density estimation to create two probabilistic maps, which are then employed to detect motion using a Markov field approach. However, it has been observed in traditional methods that reliance on the handcrafted feature makes it difficult to detect objects in underwater scenarios.

2.2 Deep Learning Based Methods

The deep learning algorithms extract the statistical distribution from training data collection. The convolution-based design leverages the spatial correlation of the image to identify the object as a region of interest [17,18,20]. To reduce the loss induced by convolution networks [10] applies dynamic aggregation and feature enhancement methods to detect the object from underwater images. Further, transformer-based architectures [2,21] can be utilized to detect the object by evaluating the attention score. Nevertheless, the shape and structure information is lost. In response to this issue, Vajpai et al. [1] introduced a U-Net architecture with an encoder-decoder structure for moving object detection. The said method still fails to detect the small object. Kapoor et al. [16] suggested that this could be due to the losses induced by the convolution kernel and information loss due to pooling. Therefore, the authors have suggested graph refactoring in a latent space to restructure the node information and preserve the spatial-contextual relationship. The said method samples the neighbourhood and uses mean, max, and LSTM aggregators to merge the information among the neighbours. As noted in the literature by [3], previous research has shown that more optimal approaches exist than utilizing a single operator for merging information. Moreover, utilizing a large amount of information from neighbouring sources retrieves more crucial information. Hence, using multiple aggregators can help propagate messages among graph nodes.

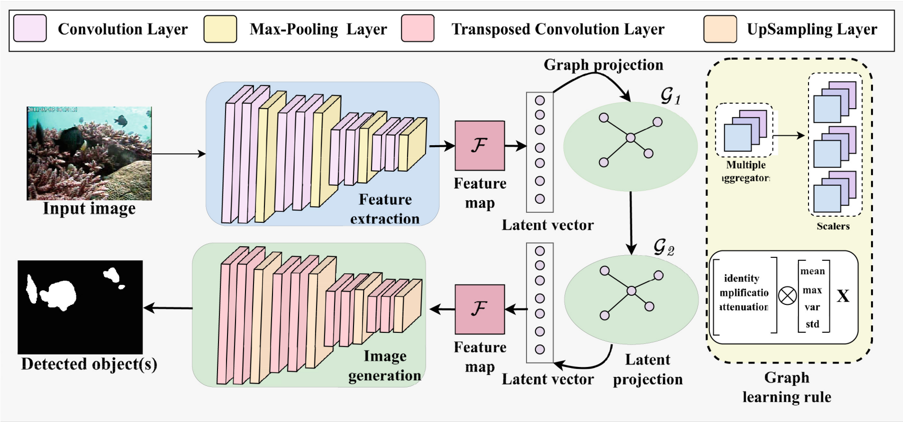


Fig. 1. Architecture of the proposed method using principal neighbourhood aggregation

3 Proposed Method

As shown in Fig 1, an encoder-decoder architecture is employed to discern the object from the background. The encoder network uses a convolution kernel to extract the features from the image. Max-pooling is employed to decrease the dimensionality of feature maps. The convolution kernel extracts the spatial relationships in Euclidean space. However, such a relationship is not possible in higher-order space. Therefore, a graph projection is implemented by assigning each element of the latent vector as a node in the graph. The information is transmitted among the adjacent nodes of the graph. Instead of using a single aggregator operator, we employed a collection of aggregator operators. After aggregating the information, the graph is projected onto feature space. The latent vector is fed into the decoder architecture to detect the moving object(s) in the given scene.

3.1 Feature Extraction

The input image is passed via the encoder module to extract features. The encoder module consists of four blocks. Every block consists of two convolutional layers, followed by a pooling layer. After traversing four blocks, the feature map is flattened to produce a latent vector. After undergoing two or three convolutions, the spatial relationship begins to weaken, resulting in the emergence of increasingly abstract features. Further, the operation of max-pooling is irreversible, i.e. the inverse operation doesn't exist. Therefore, the encoder block results in a loss of information. To mitigate this effect, graph-based learning techniques are employed.

3.2 Graph Learning Module

The graph-learning module projects the latent vector into graph space by initializing a graph and considering each element of the latent vector as a node of the graph. If the nodes share information, an edge is said to be present between them. If two nodes i and j are connected, a message is passed between them. Here, M_t maps the message being passed from one node to another by considering the hidden state of the node and the edge feature between them. After the message is passed, the vertex update function U_t will update the hidden state of the node based on the previous hidden state and message from the neighbourhood. The information is passed through the neighbourhood nodes, aggregated at the central node, and fed forward. Different aggregators, such as mean, max, and LSTM, can be used to combine the information coming from different sets of nodes. However, it is challenging to determine how to aggregate the information. Corso et al. [3] suggested using multiple sets of aggregators to transmit the information among the nodes. Hence, the principal neighbourhood aggregation method can be used to share information among the neighbours and retain maximum information. The graph learning rule as per message passing neural network is given as:

$$h_v^{t+1} = U_t(h_v^t, \sum_{\omega \in \mathcal{N}(v)} M_t(h_v^t, h_w^t, e_{vw})) \tag{1}$$

Here, h_v^t is the hidden state at a node v, U_t is the vertex update function, M_t is the message passing function at t step, e_{vw} is the edge feature. The modified message-passing function [3] with a more generalized rule can be written as,

$$h_v^{t+1} = U_t(h_v^t, \bigoplus_{\omega \in \mathcal{N}(v)} M_t(h_v^t, h_w^t, e_{vw})) \tag{2}$$

Here, $\oplus$ is a generalized aggregator function and is defined as,

$$\bigoplus = \begin{bmatrix} I \\ A \\ D \end{bmatrix} \otimes \begin{bmatrix} max_{j \in \mathcal{N}_i} X_j^l \\ mean_{j \in \mathcal{N}_i} X_j^l \\ \sigma^2(X) \\ \sigma(X) \end{bmatrix} \tag{3}$$

The scalers represent the degree of the node and can be used to amplify and attenuate the message. This is critical in a network of multiple layers. A minute change may amplify or attenuate the gradients. Hence, a number of scalers with a generalized summation operator can be utilized. In Eq. 3, the first matrix is the matrix of scalars: I represents identity, A represents amplification, and D represents attenuation. The second tensor contains a list of aggregators. In this work, we have used max, mean, variance (σ^2), and standard deviation (σ). The max and mean are simple operators and consider the max and mean values from the neighbourhood and can be defined as,

$$max_i(X^l) = max_{j \in \mathcal{N}_i} X_j^l \tag{4}$$

$$mean_i(X^l) = mean_{j \in \mathcal{N}_i} X_j^l \tag{5}$$

The variance $\sigma^2(X)$ and standard deviation σ are used to diversify the information at a node. They are defined with the expectation $\mathbb{E}$ and can be described as:

$$\sigma(X) = \sqrt{\mathbb{E}[X^2] - \mathbb{E}[X]^2} \tag{6}$$

$$\sigma^2(X) = \mathbb{E}[X^2] - \mathbb{E}[X]^2 \tag{7}$$

The node's degree (d) is used to adjust the aggregated information proportionally. A slight alteration in the degree results in exponential attenuation or amplification. Therefore, the scaling function $S(d, \alpha)$ is applied to the aggregated tensor and is defined as follows:

$$S(d, \alpha) = \left(\frac{log(d+1)}{\delta} \right)^{\alpha} \tag{8}$$

Here, α is zero for identity function (I), –1 for attenuation (D), and +1 for amplification (A). Two principal neighbourhood graph learning layers are used to refactor the node relationships and pass the message among nodes. After the message aggregation, the feature vector is built by projecting the hidden vector of the last layer by drawing a projection map from the graph to latent space.

3.3 Image Generation

The reformed latent vector is converted into image space using a decoder module. The decoder module comprises four transposed convolution modules. To maintain the semantic features, residual connections are incorporated from the encoder to the decoder. A sigmoid layer is applied at the end of the decoder, which projects the latent vector onto the image space. Binary cross entropy loss is used as a cost function to converge the model, and the error is backpropagated.

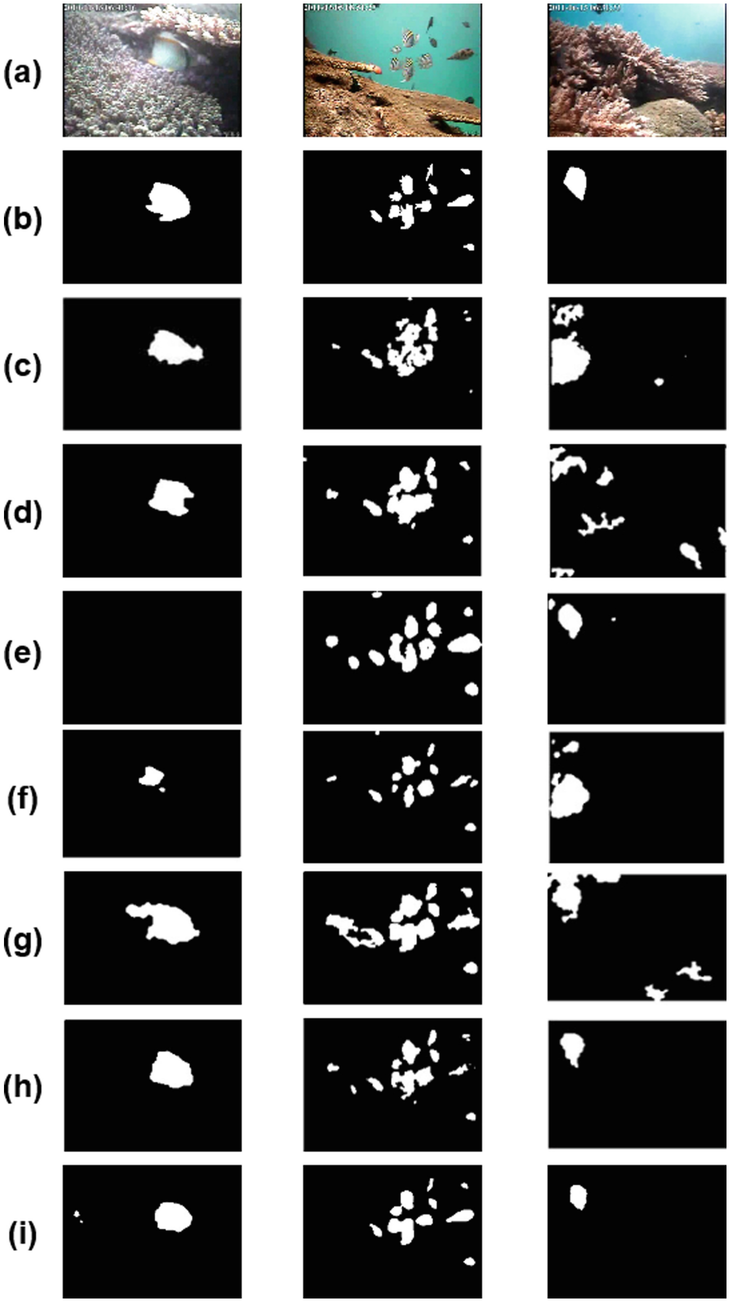


Fig. 2. Visual analysis of Fish4Knowledge dataset on the proposed method. Here, (a) original image, (b) ground-truth, (c) Texture-BGS [9], (d) ML-BGS [41], (e) MultiCueBGS [23], (f) SuBSENSEBGS [35], (g) SILTP [8], (h) MFI [38], (i) Proposed Method

4 Results and Conclusion

To assess the system's effectiveness, the suggested system is subjected to testing using all the challenges present in the Fish4Knowledge dataset and underwater change detection. The proposed technique is evaluated using twenty-four state-of-the-art algorithms: GMM [36], KDE [4], SuBSENSE [35], Vu Meter [7], Wronksian [30], ML-BGS [41], MC-BGS [23], FgSegNet [19], MFI [38], CSLTP [37], MSGAN [25], Texture-BGS [9], SILTP [8], GSMM [28], AGMM [45], ABMM [15], ADE [46], GWFT [27], HMLS [32], SSSR [13], DeColor [44], SRPCA [14], MSCL [11], GFL [39], 2PRPCA [6], OMoGMF+TV [42], CS-RPCA [12].

Table 1. Average F-Measure on four challenges of Fish4Knowledge dataset. Red indicates best, and blue indicates second best.

Method	ComplexBkg	DynamicBkg	Crowded	Hybrid	Overall
GMM [36]	0.12	0.06	0.26	0.15	0.15
ML-BGS [41]	0.58	0.32	0.74	0.46	0.52
MC-BGS [23]	0.48	0.33	0.68	0.72	0.55
KDE [4]	0.13	0.07	0.20	0.16	0.14
SuBSENSEBGS [35]	0.21	0.81	0.67	0.42	0.53
FgSegNet [19]	0.64	0.39	0.68	0.60	0.58
MFI [38]	0.83	0.64	0.69	0.64	0.70
CSLTP [37]	0.80	0.76	–	0.83	0.79
MSGAN [25]	0.92	0.79	0.74	0.89	0.83
HMLS [32]	0.84	0.90	0.84	0.91	0.87
GraphSage [16]	0.99	0.99	0.98	0.98	0.99
Proposed	0.99	0.98	0.99	0.99	0.99

4.1 Experimental Setup

We evaluate the proposed method on NVIDIA A100 80GB GPU and 256GB of RAM. The proposed method is implemented utilizing the PyTorch framework on a Linux-based operating system. To achieve faster convergence, the ResNet 50 backbone is used in the encoder, which is pre-trained with the ImageNet weights. The weights are then updated on the Fish4Knowledge dataset to incorporate the challenges posed by the underwater environment. The DGL library is employed for the primary aggregation algorithm. The model is trained using the Adam optimizer with a e^{-3} learning rate for 100 epochs. A scaling factor of 0.5 is employed, and just two layers of graphs are utilized to prevent excessive smoothing. The proposed method is evaluated on Fish4Knowledge in five challenges: complex background, crowded, dynamic background, hybrid, and standard. Further, experiments were conducted on five underwater change detection challenges: caustics, fish swarn, marine snow, small aquaculture, and two fishes.

Table 2. Quantitative analysis in terms of F-measure with thirteen SOTA architectures. The red color indicates the best, and the blue indicates the second best.

Models	Caustics	Fish Swarm	Marine Snow	small Aquaculture	two fishes	Average
ABMM [15]	0.06	0.65	0.43	0.67	0.76	0.51
AGMM [45]	0.30	0.82	0.74	0.74	0.79	0.68
GSMM [28]	0.57	0.84	0.77	0.55	0.79	0.70
ADE [46]	0.59	0.82	0.88	0.75	0.71	0.75
GWFT [27]	0.85	0.91	0.93	0.67	0.82	0.84
SSSR[13]	0.79	0.80	0.83	0.85	0.91	0.83
DeColor [44]	0.69	0.72	0.73	0.81	0.84	0.75
SRPCA [14]	0.72	0.73	0.75	0.80	0.82	0.76
MSCL [11]	0.75	0.74	0.80	0.83	0.89	0.80
GFL [39]	0.77	0.75	0.76	0.77	0.84	0.75
2PRPCA [6]	0.71	0.76	0.72	0.74	0.82	0.70
OMoGMF+TV [42]	0.66	0.70	0.70	0.67	0.79	0.70
CS-RPCA [12]	0.81	0.83	0.85	0.88	0.95	0.86
Proposed	0.99	0.99	0.99	0.98	0.99	0.99

4.2 Qualitative Analysis

To evaluate the quality of the detected objects, we conducted a visual analysis using six state-of-the-art methods: Texture-BGS [9], ML-BGS [41], MultiCueBGS [23], SuBSENSEBGS [34], SILTP [8], MFI [38], as depicted in Fig. 2. MultiCueBGS [23] fails to accurately detect the object in the complex background, as demonstrated in the first image of column e. Texture-BGS [9], ML-BGS [41], and SILTP [8] do not maintain the shape and structure of the object. Additionally, the sea weed's quasi-periodic motion also contributes to the generation of a significant amount of noise, as demonstrated in third image of Texture-BGS [9], ML-BGS [41], MultiCueBGS [23], SuBSENSEBGS [34], SILTP [8]. In second image of Texture-BGS [9], ML-BGS [41], MultiCueBGS [23], SuBSENSEBGS [34], SILTP [8], MFI [38], many rock formations are recognized as mobile entities. Our proposed technique accurately classifies the identified object as foreground while eliminating the background noise. Hence, the proposed method outperforms the state-of-the-art methods.

4.3 Quantitative Analysis

To evaluate the quantitative performance of the proposed system, an analysis is drawn in terms of the accuracy, precision, recall, and F-measure. Table 1 presents the average F-measure obtained from nine state-of-the-art methods. The proposed method is tested against thirteen state-of-the-art methods: GMM [36], KDE [4], SuBSENSE [35], Vu Meter [7], Wronksian [30], ML-BGS [41], MC-BGS [23], FgSegNet [19], MFI [38], CSLTP [37], MSGAN [25], HMLS [32], and GraphSage [16]. The proposed method outperforms the existing state-of-the-art methods. We observe a high increase of around 20% in the overall F-measure of the Fish4Knowledge dataset [5] from the learning-based method, i.e. MSGAN [25]. GraphSage [16] utilises a single aggregator across the neighbourhood, while the method proposed by us utilises four different aggregators to pass the message. The proposed method performed better in crowded and hybrid environments. Hence, in real-life underwater scenarios with multiple fishes present, our method outperforms all the existing methods. Further, we evaluated the efficiency of the proposed method on the underwater change detection net. Table 2, a quantitative analysis is drawn in terms of F-measure against thirteen state-of-the-art methods: GSMM [28], AGMM [45], ABMM [15], ADE [46], GWFT [27], SSSR [13], DeColor [44], SRPCA [14], MSCL [11], GFL [39], 2PRPCA [6], OMoGMF+TV [42], and CS-RPCA [12]. Our method surpassed all the existing methods and corroborated our findings. The quantitative analysis in terms of accuracy, precision, recall, and F-measure on all the challenges posed by the Fish4Knowledge dataset and underwater change detection are given in Fig 3. It is clearly seen that our method is robust to different challenges posed by underwater environments.

4.4 Ablation Study

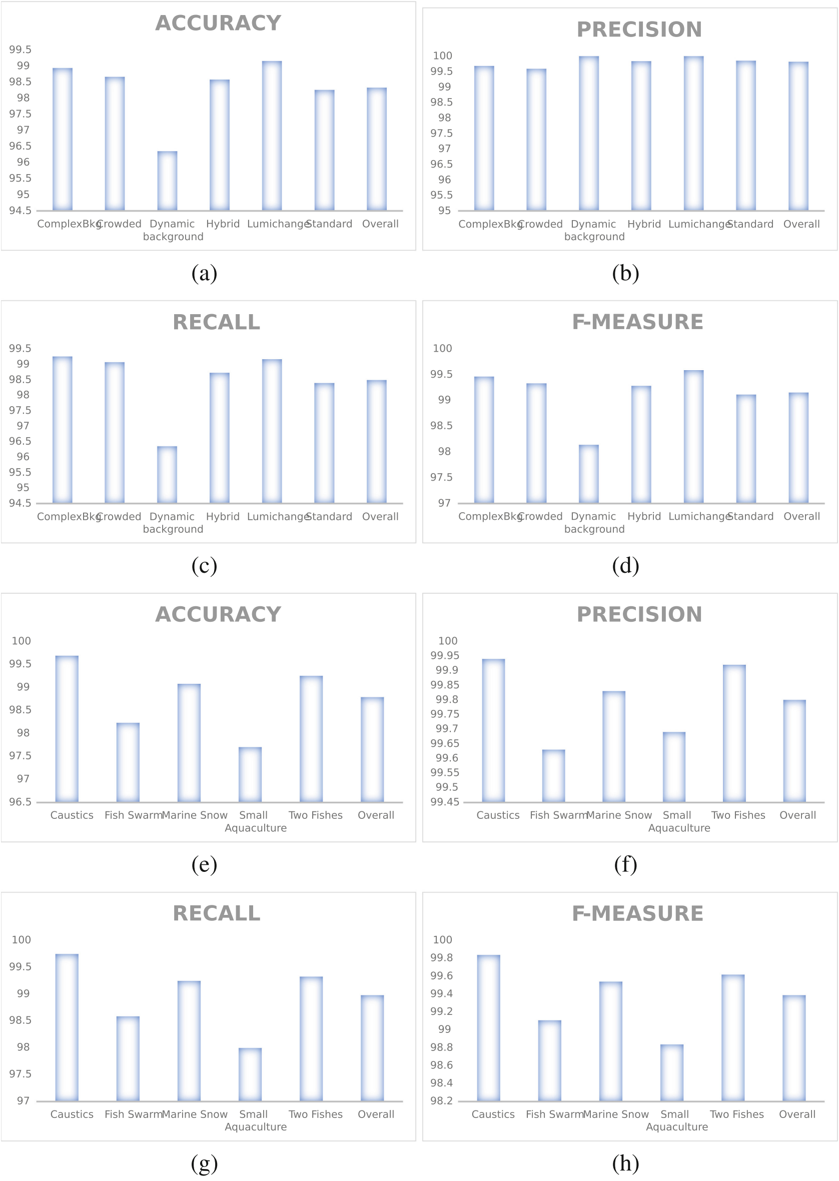
Table 3 shows the ablation study with different aggregators and scalers on complex background challenges in terms of F-measure. It shows that using different scalers and aggregators affects the performance of the proposed algorithm. The aggregator might miss information from some nodes, which introduces noise in the process. Hence, choosing a set of suitable aggregators and scalers is critical. The table shows mean, max, std, and var as aggregators and identity, amplification, and aggregator retain the most information.

Table 3. Ablation study is performed on complex background challenge of Fish4Knowledge dataset with a different set of aggregators and scalers.

Aggregator	Scalers	F-measure
mean	Identity, Amplification	99.26
mean	Identity, Amplification, attenuation	99.30
Max	Identity, Amplification	99.26
max, std	Identity, Amplification	99.41
mean, std,var	Identity,Amplification	99.13
mean, max, std, var	Identity,Amplification, attenuation	99.46

4.5 Discussion

The challenge posed by the underwater environment is more complex than that of terrestrial. Hence, there is a gap in the literature towards moving object detection in underwater scenarios. Nevertheless, the object detection problem has been addressed, but retaining the object's boundary and structure in complex underwater environments is more



(a) (b)

(c) (d)

(e) (f)

(g) (h)

Fig. 3. Quantitative analysis on all the challenges (a-d) Fish4Knowledge dataset (e-h) underwater change detection

complicated in moving object detection. Conventional deep architectures like CNNs cannot retain the object's minute details. Recently, [16] shows that information retention using different aggregators in graph space is more efficient than a CNN architecture alone. The authors showed that using mean aggregation retains most information in the graph space. However, the mean operator can increase the noise by adding information from nodes that are not required. Further, aggregators like min and max don't account for neighbourhood information. Some nodes may have slightly less information, but they may still be critical. Hence, using different aggregators can reduce the noise in the message-passing mechanism of graphs. Furthermore, different scalers are utilised to reduce the overfit and ensure the information is forwarded correctly.

5 Conclusion

This paper proposes an encoder-decoder-based architecture to detect the moving object from the background. A ResNet-50-based feature extraction module extracts spatial features from the image and projects them in the latent space. To mitigate the impact of noise, the latent vector between the encoder and decoder is projected onto a graph space. A principal neighbourhood-based refactoring of latent vectors is applied in graph space. The nodes of the projected graph undergo aggregation operations such as mean, maximum, variance, and standard deviation to derive information from their neighbouring nodes. The information is further generalized using three scalar values, identity, amplification, and attenuation, to derive a refactored set of nodes. Following refactoring, re-projection is implemented from graph space on the latent space, constructing an image using the decoder. The proposed method was compared to twenty-four state-of-the-art methods on two underwater datasets: Fish4Knowledge and underwater change detection. Our findings indicate that our proposed method outperforms all others. In the future, we will pass the message among nodes based on attention score instead of constructing the node hop neighbourhood to retain local and global information effectively.

Acknowledgement. Badri Narayan Subudhi acknowledges TIH-Drishti i-hub for providing the grant no. TIH/iHub Drishti/Project/2022-23/35.

References

1. Bajpai, V., Sharma, A., Subudhi, B.N., Veerakumar, T., Jakhetiya, V.: Underwater U-Net: deep learning with u-net for visual underwater moving object detection. In: Proceedings of the OCEANSm San Diego–Porto, pp. 1–4. IEEE (2021)
2. Chen, G., Mao, Z., Wang, K., Shen, J.: Htdet: a hybrid transformer-based approach for underwater small object detection. Remote Sensing **15**(4), 1076 (2023)
3. Corso, G., Cavalleri, L., Beaini, D., Liò, P., Veličković, P.: Principal neighbourhood aggregation for graph nets. Adv. Neural. Inf. Process. Syst. **33**, 13260–13271 (2020)
4. Elgammal, A., Duraiswami, R., Harwood, D., Davis, L.S.: Background and foreground modeling using nonparametric kernel density estimation for visual surveillance. Proc. IEEE **90**(7), 1151–1163 (2002)
5. Fisher, R.B., Chen-Burger, Y.-H., Giordano, D., Hardman, L., Lin, F.-P. (eds.): Fish4Knowledge: Collecting and Analyzing Massive Coral Reef Fish Video Data. ISRL, vol. 104. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-30208-9
6. Gao, Z., Cheong, L.F., Wang, Y.X.: Block-sparse rpca for salient motion detection. IEEE Trans. Pattern Anal. Mach. Intell. **36**(10), 1975–1987 (2014)

7. Goyat, Y., Chateau, T., Malaterre, L., Trassoudaine, L.: Vehicle trajectories evaluation by static video sensors. In: Proceedings of the IEEE Intelligent Transportation Systems Conference, pp. 864–869 (2006)
8. Han, H., Zhu, J., Liao, S., Lei, Z., Li, S.Z.: Moving object detection revisited: speed and robustness. IEEE Trans. Circuits Syst. Video Technol. **25**(6), 910–921 (2014)
9. Heikkila, M., Pietikainen, M.: A texture-based method for modeling the background and detecting moving objects. IEEE Trans. Pattern Anal. Mach. Intell. **28**(4), 657–662 (2006)
10. Hua, X., et al.: Underwater object detection algorithm based on feature enhancement and progressive dynamic aggregation strategy. Pattern Recogn. **139**, 109511 (2023)
11. Javed, S., Mahmood, A., Bouwmans, T., Jung, S.K.: Background-foreground modeling based on spatiotemporal sparse subspace clustering. IEEE Trans. Image Process. **26**(12), 5840–5854 (2017)
12. Javed, S., Mahmood, A., Dias, J., Werghi, N.: Cs-rpca: clustered sparse rpca for moving object detection. In: 2020 IEEE International Conference on Image Processing (ICIP), pp. 3209–3213. IEEE (2020)
13. Javed, S., Mahmood, A., Al-Maadeed, S., Bouwmans, T., Jung, S.K.: Moving object detection in complex scene using spatiotemporal structured-sparse rpca. IEEE Trans. Image Process. **28**(2), 1007–1022 (2018)
14. Javed, S., Mahmood, A., Bouwmans, T., Jung, S.K.: Spatiotemporal low-rank modeling for complex scene background initialization. IEEE Trans. Circuits Syst. Video Technol. **28**(6), 1315–1329 (2016)
15. KaewTraKulPong, P., Bowden, R.: An improved adaptive background mixture model for real-time tracking with shadow detection. Video-based surveillance systems: Computer vision and distributed processing, pp. 135–144 (2002)
16. Kapoor, M., Patra, S., Subudhi, B.N., Jakhetiya, V., Bansal, A.: Underwater moving object detection using an end-to-end encoder-decoder architecture and graphsage with aggregator and refactoring. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 5635–5644 (2023)
17. Li, X., Shang, M., Qin, H., Chen, L.: Fast accurate fish detection and recognition of underwater images with fast r-cnn. In: OCEANS 2015-MTS/IEEE Washington, pp. 1–5. IEEE (2015)
18. Liang, X., Song, P.: Excavating roi attention for underwater object detection. In: 2022 IEEE International Conference on Image Processing (ICIP), pp. 2651–2655. IEEE (2022)
19. Lim, L.A., Keles, H.Y.: Foreground segmentation using convolutional neural networks for multiscale feature encoding. Pattern Recogn. Lett. **112**, 256–262 (2018)
20. Liu, H., Song, P., Ding, R.: Towards domain generalization in underwater object detection. In: 2020 IEEE International Conference on Image Processing (ICIP), pp. 1971–1975. IEEE (2020)
21. Liu, Y., et al.: Dp-fishnet: dual-path pyramid vision transformer-based underwater fish detection network. Expert Syst. Appl. **238**, 122018 (2024)
22. Manderson, T., et al.: Vision-based goal-conditioned policies for underwater navigation in the presence of obstacles. arXiv preprint arXiv:2006.16235 (2020)
23. Noh, S.J., Jeon, M.: A new framework for background subtraction using multiple cues. In: Lee, K.M., Matsushita, Y., Rehg, J.M., Hu, Z. (eds.) ACCV 2012. LNCS, vol. 7726, pp. 493–506. Springer, Heidelberg (2013). https://doi.org/10.1007/978-3-642-37431-9_38
24. Palazzo, S., Kavasidis, I., Spampinato, C.: Covariance based modeling of underwater scenes for fish detection. In: 2013 IEEE International Conference on Image Processing, pp. 1481–1485 (2013). https://doi.org/10.1109/ICIP.2013.6738304
25. Patil, P.W., Thawakar, O., Dudhane, A., Murala, S.: Motion saliency based generative adversarial network for underwater moving object segmentation. In: Proceedings of the IEEE International Conference on Image Processing (ICIP), pp. 1565–1569 (2019)

26. Prabowo, M.R., Hudayani, N., Purwiyanti, S., Sulistiyanti, S.R., Setyawan, F.X.A.: A moving objects detection in underwater video using subtraction of the background model. In: 2017 4th International Conference on Electrical Engineering, Computer Science and Informatics (EECSI), pp. 1–4 (2017). 10.1109/EECSI.2017.8239148
27. Radolko, M., Farhadifard, F., von Lukas, U.: Change detection in crowded underwater scenes-via an extended gaussian switch model combined with a flux tensor pre-segmentation. In: International Conference on Computer Vision Theory and Applications, vol. 5, pp. 405–415. SCITEPRESS (2017)
28. Radolko, M., Gutzeit, E.: Video segmentation via a gaussian switch background model and higher order markov random fields. In: VISAPP (1), pp. 537–544 (2015)
29. Rout, D.K., Bhat, P.G., Veerakumar, T., Subudhi, B.N., Chaudhury, S.: A novel five-frame difference scheme for local change detection in underwater video. In: Proceedings of the Fourth International Conference on Image Information Processing (ICIIP), pp. 1–6. IEEE (2017)
30. Rout, D.K., Subudhi, B.N., Veerakumar, T., Chaudhury, S.: Spatio-contextual Gaussian mixture model for local change detection in underwater video. Expert Syst. Appl. **97**, 117–136 (2018)
31. Rout, D.K., Kapoor, M., Subudhi, B.N., Thangaraj, V., Jakhetiya, V., Bansal, A.: Underwater visual surveillance: a comprehensive survey. Ocean Eng. **309**, 118367 (2024)
32. Salman, A., et al.: Automatic fish detection in underwater videos by a deep neural network-based hybrid motion learning system. ICES J. Mar. Sci. **77**(4), 1295–1307 (2020)
33. Singh, H., Suman, S., Subudhi, B.N., Jakhetiya, V., Ghosh, A.: Action recognition in dark videos using spatio-temporal features and bidirectional encoder representations from transformers. IEEE Trans. Artifi. Intell. **4**(6), 1461–1471 (2022)
34. St-Charles, P.L., Bilodeau, G.A., Bergevin, R.: Flexible background subtraction with self-balanced local sensitivity. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, pp. 408–413 (2014)
35. St-Charles, P.L., Bilodeau, G.A., Bergevin, R.: Subsense: a universal change detection method with local adaptive sensitivity. IEEE Trans. Image Process. **24**(1), 359–373 (2014)
36. Stauffer, C., Grimson, W.E.L.: Learning patterns of activity using real-time tracking. IEEE Trans. Pattern Anal. Mach. Intell. **22**(8), 747–757 (2000)
37. Vasamsetti, S., Mittal, N., Neelapu, B.C., Sardana, H.K.: 3d local spatio-temporal ternary patterns for moving object detection in complex scenes. Cogn. Comput. **11**, 18–30 (2019)
38. Vasamsetti, S., Setia, S., Mittal, N., Sardana, H.K., Babbar, G.: Automatic underwater moving object detection using multi-feature integration framework in complex backgrounds. IET Comput. Vision **12**(6), 770–778 (2018)
39. Xin, B., Tian, Y., Wang, Y., Gao, W.: Background subtraction via generalized fused lasso foreground modeling. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4676–4684 (2015)
40. Xu, K., Hu, W., Leskovec, J., Jegelka, S.: How powerful are graph neural networks? arXiv preprint arXiv:1810.00826 (2018)
41. Yao, J., Odobez, J.M.: Multi-layer background subtraction based on color and texture. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 1–8. IEEE (2007)
42. Yong, H., Meng, D., Zuo, W., Zhang, L.: Robust online matrix factorization for dynamic background subtraction. IEEE Trans. Pattern Anal. Mach. Intell. **40**(7), 1726–1740 (2017)
43. Zaki, N., Qin, W., Krishnan, A.: Graph-based methods for cervical cancer segmentation: Advancements, limitations, and future directions. AI Open (2023)
44. Zhou, X., Yang, C., Yu, W.: Moving object detection by detecting contiguous outliers in the low-rank representation. IEEE Trans. Pattern Anal. Mach. Intell. **35**(3), 597–610 (2012)

45. Zivkovic, Z.: Improved adaptive gaussian mixture model for background subtraction. In: Proceedings of the 17th International Conference on Pattern Recognition, ICPR 2004. vol. 2, pp. 28–31. IEEE (2004)
46. Zivkovic, Z., Van Der Heijden, F.: Efficient adaptive density estimation per image pixel for the task of background subtraction. Pattern Recogn. Lett. **27**(7), 773–780 (2006)

Dynamic Loss Decay Based Robust Oriented Object Detection on Remote Sensing Images with Noisy Labels

Guozhang Liu, Ting Liu, Mengke Yuan, Tao Pang, Guangxing Yang, Hao Fu,
Tao Wang(✉), and Tongkui Liao

Piesat Information Technology, Beijing, China
{fuhao_zn,wangtao,liaotongkui}@piesat.cn

Abstract. The ambiguous appearance, tiny scale, and fine-grained classes of objects in remote sensing imagery inevitably lead to the noisy annotations in category labels of detection dataset. However, the effects and treatments of the label noises are underexplored in modern oriented remote sensing object detectors. To address this issue, we propose a robust oriented remote sensing object detection method through dynamic loss decay (DLD) mechanism, inspired by the two phase "early-learning" and "memorization" learning dynamics of deep neural networks on clean and noisy samples. To be specific, we first observe the end point of early learning phase termed as **EL**, after which the models begin to memorize the false labels that significantly degrade the detection accuracy. Secondly, under the guidance of the training indicator, the losses of each sample are ranked in descending order, and we adaptively decay the losses of the top K largest ones (bad samples) in the following epochs. Because these large losses are of high confidence to be calculated with wrong labels. Experimental results show that the method achieves excellent noise resistance performance tested on multiple public datasets such as HRSC2016 and DOTA-v1.0/v2.0 with synthetic category label noise. Our solution also has won the 2st place in the "fine-grained object detection based on sub-meter remote sensing imagery" track with noisy labels of 2023 National Big Data and Computing Intelligence Challenge.

1 Introduction

Extensive researches have been devoted to recognize objects in remote sensing imagery and locate them with more precise oriented bounding boxes, i.e. oriented remote sensing object detection (ORSOD), which is of great interest in both computer vision and remote sensing community. Most of them focus on improving performance with cutting-edge detection frameworks, such as anchor-based

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/978-3-031-78110-0_27.

A. Antonacopoulos et al. (Eds.): ICPR 2024, LNCS 15329, pp. 413–428, 2025.
https://doi.org/10.1007/978-3-031-78110-0_27

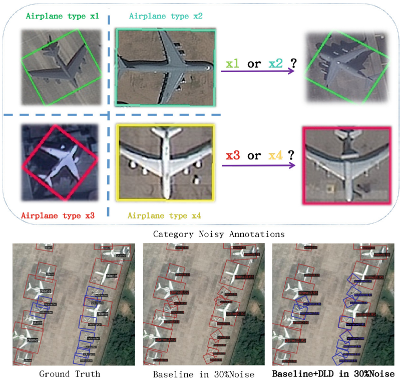


Fig. 1. The first row illustrates the difficulty of annotating the fine-grained types of planes in remote sensing images with similar appearances. The second row shows one example image with correct reference ground truth annotations in DOTA-v1.0 dataset on the left. The red and blue oriented bounding boxes indicate "plane" and "helicopter". We train a baseline ORSOD model(Oriented R-CNN) and baseline + **DLD(ours)** with synthesized 30% noisy category labels. Their detection results are visualized in the middle and right respectively, baseline result contains several false classification instances. Baseline with **DLD** generates more accurate results. (Color figure online)

one/two-stage detection [7,13,38,42], anchor-free point-based detection [11,24] and DETR-based detection [5,46]. Some other works pay attention to design network components, like augmented backbones [25,33], elaborated loss functions and angle coders [41,43–45], effective label assigners [16,39]. However, few ORSOD methods take notice of the ubiquitous and detrimental noisy annotations in current dataset, since the data set with high quality, low cost, and large scale can not be simultaneously achieved.

The noisy annotations have aroused concerns in training image classification, segmentation and object detection models for decades. A series of robust loss functions [1,9,28] and clean sample selection methods [10,12,20] are proposed to alleviate the effects of noisy labels in image classification. Like-wisely, in more challenging image segmentation task, annotation noises resistant methods [18,23,27,34] are designed to avoid the degradation. More related object detection methods [3,15,22,26,35,40,48] pay more attention to inaccurate horizontal bounding boxes or mixture of categorical and positional annotation noises. None of them addresses the category annotation noises in training ORSOD model. In practice, distinguishing the specialized type of small size, dense arranged objects (such as ships and planes) with remote sensing imagery of sub-meter spatial resolution is difficult even for the experts as shown in the first row of Fig. 1. The inter-class discrepancy is small for fine-grained plane types. It isn't surprising that the datasets contains many category label noises. In contrast to the inaccurate bounding box, the absolute wrong class label deserves particularly consideration to acquire more robust ORSOD model.

To address the challenge in training model with category noisy labels, we propose a robust oriented remote sensing object detection method through dynamic loss decay mechanism, inspired by the two phase "early-learning" and "memorization" learning dynamics of deep neural networks (DNNs) on clean and noisy samples [2,47]. Although, [47] points that DNNs can easily fit a random labeling of the training data, the quantitative differences are demonstrated in DNNs optimization on clean and noisy data [2]. The DNNs learn simple consistently shared patterns among training samples before memorizing irregular false labels. The learning dynamics of DNNs is adopted in training classification [28] and segmentation models [27] with noisy annotations. Imposing early-learning regularization [28] or correcting probable false labels [27] in early learning stage are proved to be effective. Similarly, in training ORSOD model with noisy category labels, we have observed consistent accelerated and decelerated improvement dynamics of both **mAP** (measured using model output and ground truth), and top-1 accuracy **ACC** (measured using model output and noisy annotations), which can be used to find the endpoint of early-learning phase represented by **EL**. Specifically, we first identify **EL** through monitoring the second-derivative of **ACC** curve during training. The accuracy curves of models trained on category labels with different proportion of noises share the same trend and **EL**. Secondly, the dynamic loss decay begins in the memorization phase. Since the larger loss value means the higher probability of being calculated with false labels, the overall loss computation is divided into two parts, the top K largest samples losses and the rest samples losses. We adaptively decrease the weight of the largest top K samples losses which contains the most false category labels in the following epochs. Experimental results corroborate that our method is robust to categorical annotation noises in ORSOD, and effectively decreases the model performance degradation when training on manually contaminated public dataset HRSC2016 and DOTA-v1.0/v2.0. Our solution also has won the 2st place in the "fine-grained object detection based on sub-meter remote sensing imagery" track with artificial class noise annotations of 2023 National Big Data and Computing Intelligence Challenge (NBDCIC 2023) [6]. In summary, our contributions are:

- We propose the first robust ORSOD method against categorical annotation noises through dynamic loss decay mechanism.
- We identify the effective early-learning phase endpoint **EL** in training accuracy curves through theoretical and experimental analysis for ORSOD.
- We validate the superiority of proposed method in both common ORSOD benchmarks and competitive NBDCIC 2023.

2 Related Works

2.1 Oriented Remote Sensing Object Detection

Most object detection methods commonly utilize horizontal bounding box (HBB) to localize general objects. Considering the severe overlapping by using HBB to represent bird-view remote sensing objects, oriented bounding box (OBB) representation is more accurate with extra direction. There are both similarities and discrepancies between HBB based general object detection and OBB based remote sensing object detection. On the one hand, inspired by HBB based object detection framework, representative anchor-based one-stage (R^{3}Det [42], S^{2}A-Net [13]) and two-stage (ROI-Transformer [7], Oriented-RCNN [38]), as well as anchor-free point-based (CFA [11], Oriented RepPoints [24]) and DETR-based (A^{2}O-DETR [5], ARS-DETR [46]) OBB remote sensing object detectors are proposed. On the other hand, to accommodate the additional rotation variance, backbones are augmented with rotation varied-size window attention (RVSA [33]) and large selective kernel (LSKNet [25]).

2.2 Learning with Noisy Labels

Note-worthily, the acquisition and annotation cost of remote sensing data is much higher than ground-based data. Training models with incomplete, inexact, and inaccurate supervision which collectively referred as weakly supervised in [51] is practical and desirable. Researchers are highly motivated to construct noise-resistant models in classification, segmentation, and object detection tasks.

Classification. Various tactics such as robust loss functions (GJS [9], SLC [1], ELR [28]), sample selection(MentorNet [20], co-teaching [12], OT-filter [10], CoDis [37]) and twin contrastive learning model(TCL [17]) are designed to handle categories label noises in classification. In remote sensing, [19,32,49] address the issue of noisy labels in hyperspectral image classification. Specifically, both ELR [28] and our proposed object category noise treatment are built upon the learning dynamics of deep neural networks (DNNs) on clean and noisy samples as disclosed in [47] and [2]. The DNNs have been observed to first fit clean labels during "early learning phase", and then memorize false labels in "memorization phase".

Segmentation. Image segmentation performs dense pixel-wise classification, and is more challenging. Both in the medical and remote sensing domain, the noisy annotations are ubiquitous on account of weary labeling and expertise requirement. Medical segmentation methods improve the robustness by modeling human annotation errors [18], adopting regularization term [34], etc. Note that in contrast to noisy label classification, not all semantic categories share the synchronous learning dynamics in segmentation, ADELE [27] separately correct noisy label for each category. In [23], the authors propose a semi-supervised segmentation method to handle both incomplete and inaccurate labels via multiple diverse learning groups.

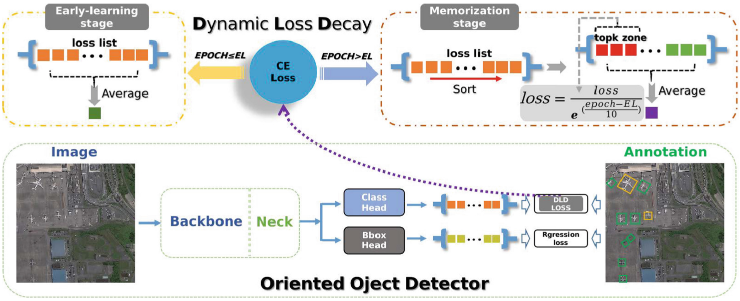


Fig. 2. The overview training process of **DLD**. The bottom part illustrates structure of an Oriented Object Detector, the upper part shows the conceptual illustration of DLD based on early-learning stage and memorization stage theory.

Object Detection. The complexity of object detection with noisy annotations lies in simultaneously dealing with possible categorical and positional noises. Efforts have been devoted to robust HBB based object detection [22,40,48], while we can hardly found studies on OBB based noise-resistant remote sensing object detection. Object detection with noisy labels is closely related to weakly-supervised object detection (WSOD) [4]. The commonly WSOD setting is train detectors with image-level labels and follows multiple instance learning (MIL) pipeline to joint optimization of object appearance and object region in positive bags.

3 Methods

3.1 Preliminary

In ELR [28], the authors theoretically interpret the learning dynamics of DNNs from gradient analysis. Considering the classification problem of training DNNs with N samples $\{\boldsymbol{x}_i, \boldsymbol{a}_i\}_{i=1}^{N}$ in C classes, where $\boldsymbol{x}_i \in \mathcal{R}^d$ is the i-th sample and $\boldsymbol{a}_i \in \{0,1\}^C$ is the corresponding one-hot annotation vector. The DNNs $\mathcal{D}_\Theta$ parameterized with Θ encodes $\boldsymbol{x}_i$ into C-dimensional feature $\mathcal{D}_\Theta(\boldsymbol{x}_i) \in \mathcal{R}^C$, and we can acquire the conditional probability prediction $\boldsymbol{p}_i = \mathcal{S}(\mathcal{D}_\Theta(\boldsymbol{x}_i))$ with softmax function $\mathcal{S}$. The cross-entropy loss (1) and gradient with respect to Θ (2) can be formulated as :

$$\mathcal{L}_{\mathrm{CE}}(\Theta) := -\frac{1}{N}\sum_{i=1}^{N}\sum_{c=1}^{C} a_i^c \log p_i^c \tag{1}$$

$$\nabla\mathcal{L}_{\mathrm{CE}}(\Theta) = \frac{1}{N}\sum_{i=1}^{N} \nabla\mathcal{D}(\Theta|\boldsymbol{x}_i)\,(\boldsymbol{p}_i - \boldsymbol{a}_i) \tag{2}$$

where $\nabla\mathcal{D}(\Theta|\boldsymbol{x}_i)$ is the Jacobian matrix of DNNs encoding for i-th sample $\boldsymbol{x}_i$ with respect to Θ. Therefore, the contribution of i-th sample x_i to the gradient of class c is $\nabla \mathcal{L}_{\mathrm{CE}}^c(\Theta) = \nabla \mathcal{D}(\Theta|\boldsymbol{x}_i)\,(p_i^c - a_i^c)$. If x_i truly belongs to class c, i.e. $a_i^c = 1$, the gradient descent will forward to the right direction $\nabla \mathcal{D}(\Theta|\boldsymbol{x}_i)$, otherwise the wrong annotation will push the update of Θ to the opposite direction.

The learning dynamics of DNNs reflects the characteristics of gradient descent with noisy labels. In early-learning stage, since correct labels are in majority and the gradient descent is well correlated with optimal direction, we can expect accelerated accuracy increment. Once the magnitude of noisy gradient dominates the update, the DNNs will memorize (overfit) the noise label, and the improvement of accuracy will decelerate.

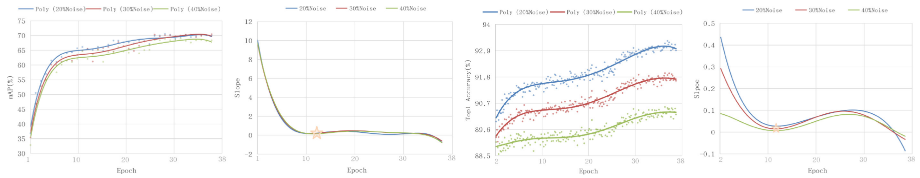


Fig. 3. The dynamics of two measurements mean average precision (**mAP**) and top 1 accuracy (**ACC**), acquired by the Oriented R-CNN with LSKNet-Tiny backbone. The experiments are conducted on DOTA-v1.0 dataset contaminated with different level of category noises (20%, 30%, and 40%). The **mAP** is calculated between model output and clean GT category labels. The **ACC** of the model output is referenced with noisy category labels. The stars in right two figures represent the early-learning endpoint.

3.2 End Point of Early-Learning in ORSOD

The existence of the turning point between early-learning phase and memorization phase, in training classification DNNs with noisy labels, has been theoretically proven in ELR [28]. The point has also been validated in the learning dynamics of image segmentation DNNs training with noisy pixel annotations in ADELE [27]. Similarly, we experimentally demonstrate that the same end point of early-learning consistently occurs in training ORSOD model with different level of noisy category labels.

Specifically, we monitor the dynamics of two measurements mean average precision (**mAP**) and top 1 accuracy (**ACC**), acquired by the representative ORSOD method Oriented R-CNN with LSKNet-Tiny backbone [25]. The experiments are conducted on the well-known DOTA-v1.0 dataset contaminated with different level of category noises (20%, 30%, and 40%), and experimental details is given in Sect. 4. The **mAP** is calculated between model output and clean ground truth (GT) category labels. In practice, we have no access to clean GT in real-world application, the model output **ACC** with reference to noisy category labels can serve as a surrogate for **mAP**.

As shown in Fig. 3, the similar trend is shared between the curves of **mAP** and **ACC**. These indexes all exhibit rapid growth in first 12 epochs and improve slower in the subsequent epochs. From the slope (first-order derivative) curve of approximated polynomial, we can identify the end epoch of the early-learning phase represented as **EL** (i.e. the initial epoch of memorization phase) by the condition that the second-derivative at **EL** is approximately equal to 0. We invite Eq. (3) to describe this condition:

$$\left|\mathbf{Poly}^{''}_{[\mathbf{ACC}_1:\mathbf{ACC}_{\mathbf{EL}}]}(\mathbf{EL})\right| < \eta \tag{3}$$

where $\mathbf{Poly}^{''}_{[\mathbf{ACC}_1:\mathbf{ACC}_{\mathbf{EL}}]}(\mathbf{EL})$ is the second derivative of the polynomial at **EL**, and the polynomial is acquired by fitting the discrete **ACC** values $[\mathbf{ACC}_1 : \mathbf{ACC}_{\mathbf{EL}}]$, and $\mathbf{ACC}_i$ stands for the **ACC** value for i-th epoch. η is a threshold and we set to 0.001 in experiments.

3.3 Dynamic Loss Decay for Robust ORSOD

Different from ELR [28] which adds early-learning regularization term in loss function and ADELE [27] which corrects high confidence false labels in early-learning stage, we design a more intuitive and effective scheme called dynamic loss decay to mitigate the influence of wrong category annotations in ORSOD. As Sect. 3.1, in the early-learning stage, the optimization is dominated by the outnumbered correct labels and will be affected by noisy labels in memorization stage. Therefore, the identified **EL** indicates the right time of intervention to avoid the detrimental influence of noisy labels in loss back propagation and misleading the ORSOD model.

The DLD mechanism is illustrated in Fig. 2 and consists of two phases divided by **EL**. In the early-learning stage, i.e. the training epoch is smaller than **EL**, we use the standard Cross Entropy loss. In the memorization stage, we select the top K samples of which the losses are the top K largest ones (more probably noisy labels), and epoch-wisely decay the losses of the top K samples as they are most probably calculated with wrong category labels. All the training samples are denoted by $\boldsymbol{X}$, the top K samples with top K largest losses are represented as $\boldsymbol{X}_K$, and $\boldsymbol{X}_r$ stands for the rest samples. The formulation of $\mathcal{L}_{\mathrm{DLD}}$ can be given in (4):

$$\mathcal{L}_{\mathrm{DLD}} = \begin{cases} \mathcal{L}_{\mathrm{CE}}(\boldsymbol{X}), & \text{if EC} < \mathbf{EL} \\ \alpha\mathcal{L}_{\mathrm{CE}}(\boldsymbol{X}_k) + \mathcal{L}_{\mathrm{CE}}(\boldsymbol{X}_r), & \text{if EC} \geq \mathbf{EL} \end{cases} \tag{4}$$

where $\alpha = \exp(\frac{10}{\mathrm{EC}-\mathbf{EL}})$ is the dynamic decay factor, $\mathcal{L}_{\mathrm{CE}}(\boldsymbol{X})$ represents the Cross Entropy loss with respect to samples $\boldsymbol{X}$, EC stands for the current epoch number. In the memorization stage, the loss function consists of two parts: the first part gradually diminishes as the number of epochs increases, and the second part remains unchanged.

4 Experiments

In this section, we first briefly introduce the testing data sets HRSC2016 [30], DOTA-v1.0/v2.0 [8,36]. Then, we describe the implementation details of the experiments. Thirdly, we report the ablation study of the key components and

Table 1. Ablation study of early-learning end point EL. We report the best **mAP** values acquired by varying the beginning epoch of the second stage of DLD on validation sets of DOTA-v1.0. The first row term "20%-LSK Tiny (12)" represents that the Oriented R-CNN model with LSK-Tiny backbone trained on DOTA-v1.0 containing 20% category label noises, and the **EL** identified by our method is 12. We can observe that the identified **EL** leads to superior robustness on different datasets and backbones.

EL	20%-LSK	30%-LSK	40%-LSK	20%-LSK	30%-LSK	40%-LSK
	Tiny(**12**)	Tiny(**12**)	Tiny(**12**)	Small(**14**)	Small(**14**)	Small(**14**)
baseline	70.5	70.3	69.0	73.2	70.7	70.2
EL-8	70.7	70.8	70.3	72.8	71.6	70.4
EL-4	71.2	**72.3**	69.1	72.6	71.6	70.6
EL	**71.9**	71.6	**70.4**	**73.4**	**72.2**	**71.0**
EL+4	71.4	71.6	69.5	73.0	70.7	70.3
EL+8	71.5	71.5	69.4	73.4	71.6	69.6

Table 2. Comparison of **mAP**(%) for different ORSOD models training with DLD. The results of DLD are highlighted. We report the best **mAP** value for validation set of DOTA-v1.0.

Dataset	Detector	0% noise	20% noise		30% noise		40% noise	
		baseline	baseline	DLD	baseline	DLD	baseline	DLD
DOTA-v1.0	Oriented R-CNN	74.2	70.5	**71.9(+1.4)**	70.3	**71.6(+1.3)**	69.0	**70.4(+1.4)**
	ROI-Transformer	74.6	71.5	**71.6(+0.1)**	70.4	**71.6(+1.2)**	68.1	**70.1(+2.0)**
	ReDet	73.2	70.5	**70.5(+0.0)**	69.9	**70.3(+0.4)**	67.7	**68.3(+0.6)**

Table 3. Comparison of **mAP**(%) for the hyperparameter **Top-K** with different proportions of category incorrect labels in the DOTA-v1.0 dataset. The selected model is the Oriented R-CNN detector with LSKNet-Tiny backbone.

Method	baseline	Top-K = 3%	Top-K = 5%	Top-K = 7%	Top-K = 10%
LSK-T-20%	70.5	71.5	**71.9**	70.5	70.8
LSK-T-30%	70.3	71.0	**71.6**	71.0	70.9
LSK-T-40%	69.0	68.6	68.9	**70.4**	70.1

in-depth analysis of proposed method to verify their effectiveness. Finally, we show the performance of proposed method compared with other competitors on both synthesized noisy ORSOD dataset and 2023 National Big Data and Computing Intelligence Challenge.

4.1 Datasets

HRSC2016 [30] is a popular ship detection dataset that contains 1,070 images and 2,976 instances using satellite imagery. It has a three level category hierarchy, and we chose to use its first and second tier level which contains four categories: aircraft carrier, warship, merchant ship and other generic ships.

DOTA-v1.0 [36] is a large scale aerial images dataset for object detection. It is widely used in develop and evaluate ORSOD methods. The dataset contains 15 categories, 2,806 images and more than 180,000 instances. The size of images varies from 800 × 800 to 20000 × 20000.

DOTA-v2.0 [8] collects more sub-meter remote sensing and aerial images. DOTA-v2.0 has 18 categories, 11,268 images and 1,793,658 instances. In our ablation study, we have assessed 17 out of 18 categories in the validation set of DOTA-v2.0, excluding the "helipad" category. The validation set comprises a total of 130,909 instances, with only 3 instances belonging to the "helipad" category in our processed dataset. The rare presence of "helipad" instances significantly degrades the **mAP** and **ACC** indexes and distorts the overall curve trends.

4.2 Implementation

We adopt single-scale training and testing strategy by cropping all images into 1024 × 1024 patches with overlap of 200 pixels. For noisy category label generation, we randomly select a proportion of instances in the annotations, and set their categories to random new ones without changing the bounding box. Oriented R-CNN [38] on MMRotate [50] framework, with LSKNet [25]-Tiny, LSKNet-Small and SwinTransformer [29]-Tiny as backbones, is adopted for different experiments respectively. NVIDIA A40 GPU is utilized to carry out all the experiments. The models have been trained for 36 epochs with AdamW optimizer. We first train the baseline model on the original dataset with clean labels. Then, we train ORSOD models with different level of noisy labels (20%, 30%, 40%) for comparisons.

4.3 Ablation Study and Analysis

In this section, we report the results of ablation study on validation sets of DOTA-v1.0 and DOTA-v2.0 to validate the effectiveness of our method.

End Point of Early-Learning. Identifying the endpoint of early learning is a crucial component of our method. We adhere to the approach outlined in Sect. 3.2 and compute **EL** using Eq. (3). The influence of changing the initial epoch of the second stage of DLD is explored in Table 1. The candidate epochs $[\mathbf{EL}-8, \mathbf{EL}-4, \mathbf{EL}+4, \mathbf{EL}+8]$ are selected around the **EL** found by our proposed criterion. In order to verify the generality of the identified **EL**, we report the best DOTA-v1.0 validation set **mAP** of different backbones (LSK-Tiny and LSK-Small) training with different noisy-level labels (20%, 30%, and 40%). For example, the head row term "20%-LSK Tiny (12)" represents the experimental setting, that the Oriented R-CNN model with LSK-Tiny backbone trained on DOTA-v1.0 containing 20% category label noises with DLD, and the

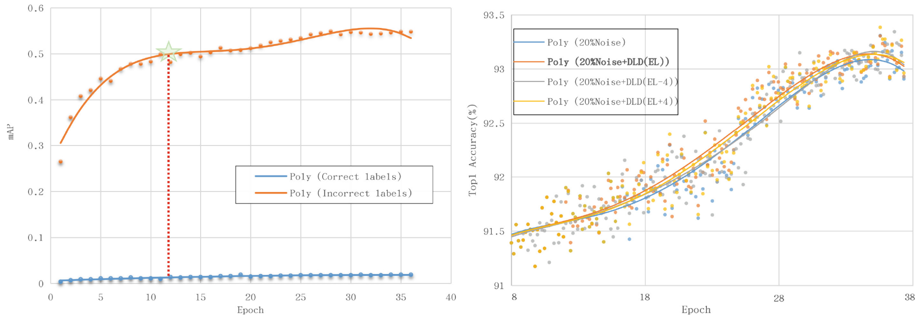


Fig. 4. Left: The more elaborated curves of **mAPC** (**mAP** with respect to correct category labels) and **mAPI** (**mAP** with respect to incorrect category labels) of training set. The Orient R-CNN model is trained with labels with 40% noise level. The curve of **mAPI** stays below 2% while **mAPC** continuously improves during whole process. **Right**: The **ACC** curves of ORSOD model training with four different strategies and labels with 20% noise. The four training strategies are training Oriented R-CNN not using DLD, using DLD and the loss decay begins at epoch EL-4 (8), EL (12), and EL+4 (16).

corresponding **EL** identified by our method is 12. From the third row to the seventh row of the second column, the initial epoch of loss decay in DLD is varied from 4 (EL-8) to 20 (EL+8).

We can observe that the **EL** selected by our proposed criterion consistently lead to superior robustness with different noise-level datasets and backbones. Although the endpoint **EL** identified by the method in Sect. 3.2 serves as a effective indicator for DLD, it's important to note that the best performance may not achieved exactly at **EL**, but in the neighbourhood of **EL** as shown in the third column of Table 1. Therefore, comparing with the baseline model, we can claim that under the guidance of **EL**, DLD can significantly boost its the noise resistance.

To further analysis the relationship between **EL** and mAP accuracy of incorrect labels, we show the more elaborated curves of **mAPC** (**mAP** with respect to correct category labels) and **mAPI** (**mAP** with respect to incorrect category labels) in training set in Fig. 4 left figure. The ORSOD baseline model is trained using labels with 40% noise. **mAPC** continuously improves, and slow down its growth rate at epoch 12 which is align with **mAP** performance in validation set as is shown in Fig. 3. This is phenomena also demonstrates that **ACC** can reflect the overall training accuracy of correct labels when model train with noisy labels. Owing to the erroneous labels being randomly reassigned across the remaining eleven categories, the **mAPI** consistently remains below 2% throughout the entire process, without any marked spikes in performance.

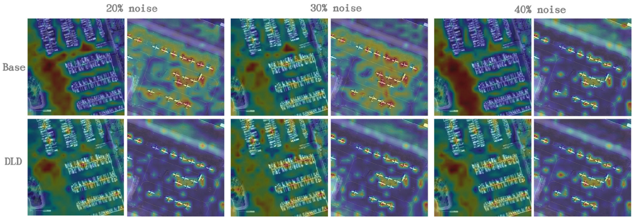


Fig. 5. The heatmap shows the distribution of region of interest with the attention map during inference, where the attention increases from blue to red. Notably, the ORSOD model is more focused on the target objects after DLD is applied. (Color figure online)

The Effectiveness of DLD. We demonstrate that the effectiveness of DLD from three aspects: analysis **ACC** curves, class activation map visualization and integration with different ORSOD methods.

Firstly, we compare the **ACC** curves of ORSOD model training with four different strategies and labels with 20% noise in Fig. 4 right figure. The four training strategies are not using DLD, using DLD and the loss decay begins at epoch **EL**-4 (8), **EL** (12), and **EL**+4 (16). The three training strategies with DLD consistently improve the **ACC**, which shows that DLD take effects at relatively relaxed **EL** neighbourhood.

Secondly, we compare the class activation maps of ORSOD model with and without DLD on test images. Some examples of EigenCAM [31] are shown in Fig. 5. Training with DLD, the ORSOD model tend to focus more on foreground targets which is beneficial for noise resistance.

Finally, to further analysis the generality of our method for different oriented object detectors, we choose two representative methods ROI-Transformer and ReDet [14], and employ LSKNet-Tiny as backbone. Meanwhile, we use DOTA-v1.0 dataset with different proportion of noise for training. The results are illustrated in Table 2 which demonstrates that our method can be easily integrated in different ORSOD models and improve their robustness.

Top-K Selection. The effectiveness of DLD heavily relies on the hyperparameter K in selecting the top K samples with largest losses. We have conduct extensive experiments to explore the impact of different K values, specifically top 3%, top 5%, top 7%, and top 10%. As illustrated in Table 3, when the incorrect label proportions are 20% and 30%, top 5% yields the highest **mAP**, while top 7% performs better in the case of 40% noise. These experimental findings conform with the intuitive expectation that dataset with larger proportion of incorrect labels should employ larger K value to decay the weight of more losses terms.

Table 4. Comparison between baseline and DLD method on three datasets: DOTA-v1.0, DOTA-v2.0, and HRSC2016. These datasets have been deliberately contaminated with category noise levels of 20%, 30%, and 40%. The employed detector employed is Oriented R-CNN, and three distinct backbones are selected: LSKNet-Tiny, LSKNet-Small, and SwinTransformer-Tiny. The reported values represent the highest accuracy from the best epoch evaluated in the validation set. DLD represents the model that is trained by our method.

Dataset	Method	0% noise	20% noise		30% noise		40% noise	
		baseline	baseline	DLD	baseline	DLD	baseline	DLD
DOTA-v1.0	LSKNet-Tiny	74.2	70.5	**71.9(+1.4)**	70.3	**71.6(+1.3)**	69.0	**70.4(+1.4)**
	LSKNet-Small	75.6	73.2	**73.4(+0.2)**	70.7	**72.2(+1.5)**	70.2	**71.0(+0.8)**
	SwinTransformer-Tiny	75.3	71.5	**71.7(+0.2)**	71.4	**72.0(+0.6)**	69.5	**70.8(+1.3)**
DOTA-v2.0	LSKNet-Tiny	66.1	63.5	**61.5(+1.4)**	63.1	**63.7(+0.6)**	62.4	**62.8(+0.4)**
	LSKNet-Small	66.4	63.1	**64.2(+1.1)**	62.9	**64.3(+1.4)**	62.2	**63.2(+1.0)**
	SwinTransformer-Tiny	67.2	63.8	**64.4(+0.6)**	62.5	**62.9(+0.4)**	61.3	**62.8(+1.5)**
HRSC2016	LSKNet-Tiny	82.1	81.1	**81.3(+0.2)**	78.6	**79.2(+0.6)**	72.7	**72.9(+0.2)**
	LSKNet-Small	86.8	85.3	**85.6(+0.3)**	83.0	**84.7(+1.7)**	80.1	**81.7(+1.6)**
	SwinTransformer-Tiny	75.3	71.5	**71.7(+0.2)**	71.4	**72.0(+0.6)**	69.5	**70.8(+1.3)**

4.4 Final Results

Test on Open Datasets. Based on Oriented R-CNN framework, we evaluate DLD with backbones of three different size (CNN based LSKNet-Tiny: 4.3M [25], CNN-based LSKNet-Small: 14.4M [25], and Transformer based SwinTransformer-Tiny: 29M [29]), three ORSOD datasets (DOTA-v1.0, DOTA-v2.0, and HRSC2016), three proportion of incorrect category labels(20%, 30%, and 40%). The results presented in Table 4 demonstrates that DLD can effectively alleviate the models' degradation caused by noisy category labels, the maximum improvement in **mAP** is 2.0% for ROI-Transformer. Meanwhile, as is shown in Table 3, endpoint of early-learning for Oriented Object Detectors can be identified effectively through our method.

DLD vs. Label Smoothing Method [21]. Label Smoothing (LS) is a regularization method involves penalizing the distribution of the network's outputs, thereby encouraging the model to be more cautious in its predictions. As shown in Table 5, DLD outperforms LS by a considerable margin in the experiments with 20% and 30% category incorrect labels, and achieves an equal mAP in the 40% noise ratio setting.

NBDCIC 2023. The competition has posed several significant challenges such as weak target features, subtle differences between classes, labeling noise, and an extremely unbalanced distribution of categories. The most noteworthy characteristic of the dataset is the category labels are randomly contaminated with

Table 5. Comparison of mAP(%) between Label Smoothing (LS) method and our method (DLD).

Method	baseline	LS	**DLD**
LSK-T-20%	70.5	71.5	**71.9**
LSK-T-30%	70.3	70.9	**72.3**
LSK-T-40%	69.0	**70.4**	**70.4**

noise and the ratio of label noise is different in the preliminary round and the final round. The competition pay close attention on the way of alleviating the degradation of model caused by category noise. Despite these challenges, our team has acquired remarkable results, as highlighted in Table 6 showcasing the superiority of proposed method in tackling label noise.

Table 6. Results of *the Fine-grained Object Detection Based on Sub-meter Remote Sensing Images* from NBDCIC 2023.

Team Name	Final Round	First Round
GaoKongTanCe Team	0.7758	0.7533
Sensing_earth(**ours**)	0.7758	0.7518
CAE_AI	0.7565	0.6994
Happy Children's Day	0.7508	0.7656
default13253462	0.7448	0.6803

5 Conclusions

In this paper, we propose the first robust oriented object detection method for remote sensing images, which addresses the issue of noisy category training labels with dynamic loss decay mechanism. Based on the theoretical analysis and experimental validation, we identify the existence of the early-learning phase and memorization phase in training ORSOD model with noisy labels, and propose a feasible approach to find the end point of early-learning **EL**. Accordingly, we design an effective dynamic loss decay scheme by gradually reduce the top K largest loss terms which are most likely calculated with false labels in subsequent epochs of **EL**. Experiments on both synthesized noisy ORSOD datasets and NBDCIC 2023 demonstrate the effectiveness of proposed DLD in preventing training category noise, and the ablation studies corroborate the rationality of the selected EL and the loss decay scheme.

References

1. Ahn, C., Kim, K., Baek, J.W., Lim, J., Han, S.: Sample-wise label confidence incorporation for learning with noisy labels. In: ICCV, pp. 1823–1832 (2023)
2. Arpit, D., et al.: A closer look at memorization in deep networks. In: ICML, pp. 233–242. PMLR (2017)
3. Bernhard, M., Schubert, M.: Correcting imprecise object locations for training object detectors in remote sensing applications. Remote Sens. **13**(24), 4962 (2021)
4. Bilen, H., Vedaldi, A.: Weakly supervised deep detection networks. In: CVPR, pp. 2846–2854 (2016)
5. Dai, L., Liu, H., Tang, H., Wu, Z., Song, P.: Ao2-detr: arbitrary-oriented object detection transformer (2022)
6. Data, T.B., Lab, D.: 2023 National Big Data and Computing Intelligence Challenge (2023). https://www.datafountain.cn/competitions/637. Accessed 17 Nov 2023
7. Ding, J., Xue, N., Long, Y., Xia, G.S., Lu, Q.: Learning roi transformer for oriented object detection in aerial images. In: CVPR, pp. 2849–2858 (2019)
8. Ding, J., et al.: Object detection in aerial images: a large-scale benchmark and challenges. IEEE TPAMI (2021). https://doi.org/10.1109/TPAMI.2021.3117983
9. Englesson, E., Azizpour, H.: Generalized jensen-shannon divergence loss for learning with noisy labels, vol. 34, pp. 30284–30297 (2021)
10. Feng, C., Ren, Y., Xie, X.: Ot-filter: an optimal transport filter for learning with noisy labels. In: CVPR, pp. 16164–16174 (2023)
11. Guo, Z., Liu, C., Zhang, X., Jiao, J., Ji, X., Ye, Q.: Beyond bounding-box: convex-hull feature adaptation for oriented and densely packed object detection. In: CVPR, pp. 8792–8801 (2021)
12. Han, B., et al.: Co-teaching: robust training of deep neural networks with extremely noisy labels, vol. 31 (2018)
13. Han, J., Ding, J., Li, J., Xia, G.S.: Align deep features for oriented object detection, vol. 60, pp. 1–11 (2021)
14. Han, J., Ding, J., Xue, N., Xia, G.S.: Redet: a rotation-equivariant detector for aerial object detection. In: CVPR, pp. 5–7, 14 (2021)
15. Hu, Z., Gao, K., Zhang, X., Wang, J., Wang, H., Han, J.: Probability differential-based class label noise purification for object detection in aerial images, vol. 19, pp. 1–5 (2022)
16. Huang, Z., Li, W., Xia, X.G., Tao, R.: A general gaussian heatmap label assignment for arbitrary-oriented object detection. IEEE TIP **31**, 1895–1910 (2022)
17. Huang, Z., Zhang, J., Shan, H.: Twin contrastive learning with noisy labels. In: CVPR (2023)
18. Jacob, J., Ciccarelli, O., Barkhof, F., Alexander, D.C.: Disentangling human error from the ground truth in segmentation of medical images. ACL (2021)
19. Jiang, J., Ma, J., Liu, X.: Multilayer spectral-spatial graphs for label noisy robust hyperspectral image classification, vol. 33, no. 2, pp. 839–852 (2020)
20. Jiang, L., Zhou, Z., Leung, T., Li, L.J., Fei-Fei, L.: Mentornet: learning data-driven curriculum for very deep neural networks on corrupted labels. In: ICML, pp. 2304–2313. PMLR (2018)
21. Krizhevsky, A., Sutskever, I., Hinton, G.: Regularizing neural networks by penalizing confident output distributions. In: ICML (2012)
22. Li, J., Xiong, C., Socher, R., Hoi, S.: Towards noise-resistant object detection with noisy annotations. arXiv preprint arXiv:2003.01285 (2020)

23. Li, P., et al.: Semi-supervised semantic segmentation under label noise via diverse learning groups. In: ICCV, pp. 1229–1238 (2023)
24. Li, W., Chen, Y., Hu, K., Zhu, J.: Oriented reppoints for aerial object detection. In: CVPR, pp. 1829–1838 (2022)
25. Li, Y., Hou, Q., Zheng, Z., Cheng, M.M., Yang, J., Li, X.: Large selective kernel network for remote sensing object detection. In: ICCV, pp. 16794–16805 (2023)
26. Liu, C., Wang, K., Lu, H., Cao, Z., Zhang, Z.: Robust object detection with inaccurate bounding boxes. In: ECCV, pp. 53–69. Springer, Heidelberg (2022)
27. Liu, S., Liu, K., Zhu, W., Shen, Y., Fernandez-Granda, C.: Adaptive early-learning correction for segmentation from noisy annotations. In: CVPR, pp. 2606–2616 (2022)
28. Liu, S., Niles-Weed, J., Razavian, N., Fernandez-Granda, C.: Early-learning regularization prevents memorization of noisy labels, vol. 33, pp. 20331–20342 (2020)
29. Liu, Z., et al.: Swin transformer: hierarchical vision transformer using shifted windows. In: ICCV, pp. 10012–10022 (2021)
30. Liu, Z., Yuan, L., Weng, L., Yang, Y.: A high resolution optical satellite image dataset for ship recognition and some new baselines. In: International Conference on Pattern Recognition Applications and Methods (2017). https://api.semanticscholar.org/CorpusID:32179046
31. Muhammad, M.B., Yeasin, M.: Eigen-cam: class activation map using principal components. In: 2020 International Joint Conference on Neural Networks (IJCNN), pp. 1–7 (2020). https://doi.org/10.1109/IJCNN48605.2020.9206626
32. Tu, B., Zhou, C., He, D., Huang, S., Plaza, A.: Hyperspectral classification with noisy label detection via superpixel-to-pixel weighting distance, vol. 58, no. 6, pp. 4116–4131 (2020)
33. Wang, D., et al.: Advancing plain vision transformer toward remote sensing foundation model, vol. 61, pp. 1–15 (2023). https://doi.org/10.1109/TGRS.2022.3222818
34. Wang, G., et al.: A noise-robust framework for automatic segmentation of covid-19 pneumonia lesions from ct images, vol. 39, no. 8, pp. 2653–2663 (2020)
35. Wu, D., Chen, P., Yu, X., Li, G., Han, Z., Jiao, J.: Spatial self-distillation for object detection with inaccurate bounding boxes. In: ICCV, pp. 6855–6865 (2023)
36. Xia, G.S., et al.: Dota: a large-scale dataset for object detection in aerial images. In: CVPR (2018)
37. Xia, X., et al.: Combating noisy labels with sample selection by mining high-discrepancy examples. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 1833–1843 (2023)
38. Xie, X., Cheng, G., Wang, J., Yao, X., Han, J.: Oriented r-cnn for object detection. In: ICCV, pp. 3520–3529 (2021)
39. Xu, C., et al.: Dynamic coarse-to-fine learning for oriented tiny object detection. In: CVPR, pp. 7318–7328 (2023)
40. Xu, Y., Zhu, L., Yang, Y., Wu, F.: Training robust object detectors from noisy category labels and imprecise bounding boxes. IEEE TIP **30**, 5782–5792 (2021)
41. Yang, X., Hou, L., Zhou, Y., Wang, W., Yan, J.: Dense label encoding for boundary discontinuity free rotation detection. In: CVPR, pp. 15819–15829 (2021)
42. Yang, X., Yan, J., Feng, Z., He, T.: R3det: refined single-stage detector with feature refinement for rotating object. In: AAAI, vol. 35, pp. 3163–3171 (2021)
43. Yang, X., Yan, J., Ming, Q., Wang, W., Zhang, X., Tian, Q.: Rethinking rotated object detection with gaussian wasserstein distance loss. In: ICML, pp. 11830–11841. PMLR (2021)
44. Yang, X., et al.: Learning high-precision bounding box for rotated object detection via kullback-leibler divergence, vol. 34, pp. 18381–18394 (2021)

45. Yu, Y., Da, F.: Phase-shifting coder: predicting accurate orientation in oriented object detection. In: CVPR, pp. 13354–13363 (2023)
46. Zeng, Y., Yang, X., Li, Q., Chen, Y., Yan, J.: Ars-detr: aspect ratio sensitive oriented object detection with transformer. arXiv preprint arXiv:2303.04989 (2023)
47. Zhang, C., Bengio, S., Hardt, M., Recht, B., Vinyals, O.: Understanding deep learning (still) requires rethinking generalization. Commun. ACM **64**(3), 107–115 (2021)
48. Zhang, X., Yang, Y., Feng, J.: Learning to localize objects with noisy labeled instances. In: AAAI, vol. 33, pp. 9219–9226 (2019)
49. Zhang, X., Yang, S., Feng, Z., Song, L., Wei, Y., Jiao, L.: Triple contrastive representation learning for hyperspectral image classification with noisy labels (2023)
50. Zhou, Y., et al.: Mmrotate. In: ACMMM. ACM (2022)
51. Zhou, Z.H.: A brief introduction to weakly supervised learning, vol. 5, no. 1, pp. 44–53 (2018)

Environment-Independent Fusion for Robust Object Detection in Adverse Environments

Wenlong Zhong[1], Yunfei Zhang[1](✉), and Si Wu[2]

[1] School of Future Technology, South China University of Technology,
Guangzhou, China
{202221062428,cszhangyunfei}@mail.scut.edu.cn
[2] School of Computer Science and Engineering, South China University of
Technology, Guangzhou, China
cswusi@scut.edu.cn

Abstract. Object detection has achieved remarkable progress, yet its efficacy undergoes substantial deterioration in challenging or adverse environmental conditions. Current domain adaptation object detection (DAOD) methodologies predominantly concentrate on transfer models to acclimate to the target domain, but neglecting the potential erosion in detection performance within the source domains. This narrow focus can undermine the comprehensive robustness and adaptability of the detection systems. We propose a simple but efficient method called Environment-Independent Fusion YOLO (EIF-YOLO) to tackle this issue. Our method focuses on extracting and fusing environment-independent features to enable accurate detection across different domains. We have reused the original feature extractor while preserving all its parameters and optimizing it by mixing data from the source and target domains. To encourage the extraction of object-related features, we introduce multi-layer perceptual regularization to align the features from the original feature extractor. Additionally, we introduce a domain-adaptive fusion that merges features from different domains while minimizing interference with the original data features. Experimental results show that our method surpasses existing foggy and low-light detection approaches while maintaining excellent source domain performance.

Keywords: Object Detection · Domain adaptation · Environment-Independent Fusion

1 Introduction

Object detection is a fundamental and vital task in computer vision, which identifies and locates specific objects in images or videos. The continuous improve-

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/978-3-031-78110-0_28.

A. Antonacopoulos et al. (Eds.): ICPR 2024, LNCS 15329, pp. 429–444, 2025.
https://doi.org/10.1007/978-3-031-78110-0_28

ment of object detection can be attributed to the emergence of efficient networks like SDD [3] and DETR [7] also the iterative optimization of detection models such as Yolo series [1,34]. These efficient methods find practical applications in various domains, including autonomous driving and security surveillance [8–10]. However, real-world scenarios often present adverse environments, including foggy days, low-light, and other factors that degrade the quality of captured images [16–18]. The domain gap between test images and training images can result in a significant performance decline of the trained model [6].

Directly training object detection for adverse environments is not feasible due to the high cost of data collection and accurate data labeling. Domain adaptation offers a solution to transfer knowledge learned from the source domain to the target domain [4,5]. One approach is aligning the target domain data with the source domain data by using the data alignment algorithms [16–18,20], which allows models trained on the source domain to perform well on the target domain. Another method involves enhancing the robustness of the feature extractor to improve the adaptive capability [25]. Additionally, constructing a student-teacher model for adjusting the detection ability of the model can be beneficial [27].

Existing DAOD methods often focus on adapting the object detection to the target domain data but overlook maintaining the detection performance on the source domain. To address this issue, we propose to leverage the capabilities of YOLOv8 by freezing its parameters and reusing its backbone as a new trainable adapter. This adapter is responsible for detecting objects in adverse environments and suppressing irrelevant object information. This approach allows the model to perform excellently in different domains, as illustrated in Fig. 1.

More specifically, we propose an Environment-Independent Fusion YOLO for robust object detection. To maintain the performance on the source domain. We reused its backbone as a learnable environment-independent adapter (EIA) for extracting features from different domains. To guide the feature extraction, we introduce a multi-layer perceptual regularization (MPR) that encourages the adapter to focus on object-related information. We employ a zero-conv [31] based domain-adaptive fusion (DAF) to enhance object detection performance. This module facilitates the adaptive fusion of features from different extractors while preserving the original capability of the fixed module. The framework of our method is depicted in Fig. 2. Experimental results on public object detection benchmarks demonstrate that our proposed structure enables us to obtain robust and efficient object detection.

The contributions are summarized as follows: 1) We use pre-trained YOLOv8 to initialize an adapter responsible for feature extraction from different domains, and we introduce multi-layer perceptual loss to motivate the adapter to focus on object-related features. 2) To prevent the performance of source object detection from interfering, we use domain-adaptive fusion to fuse features from different modules. 3) Our model outperforms competing methods on multiple publicly detection datasets while retaining the performance on the source domain.

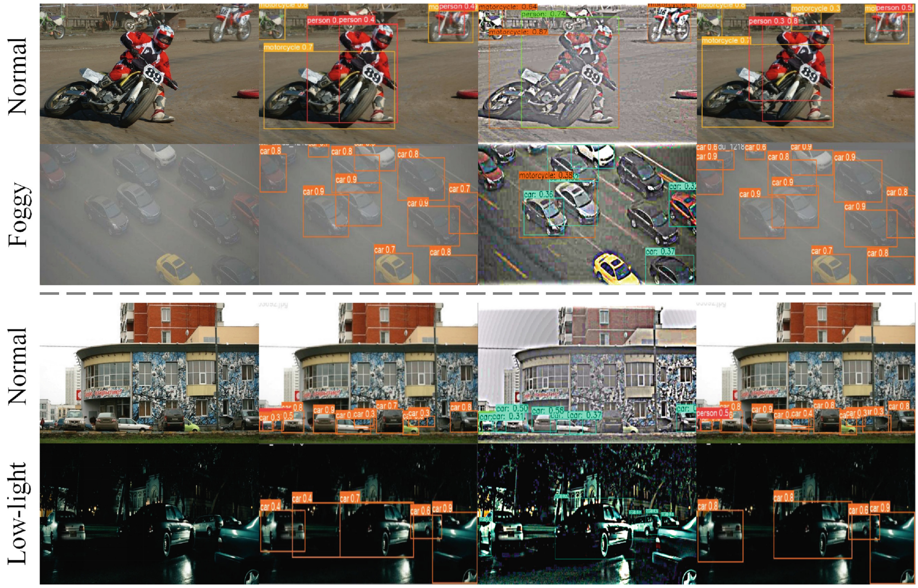


Fig. 1. Original image(Column 1) and detection results. EIF-YOLO(Column 4) demonstrates the capability to extract environment-independent features for detection across various environments, which is not achievable by the baseline(YOLOv8) method(Column 2) or the data-enhanced IA-YOLO(Column 3).

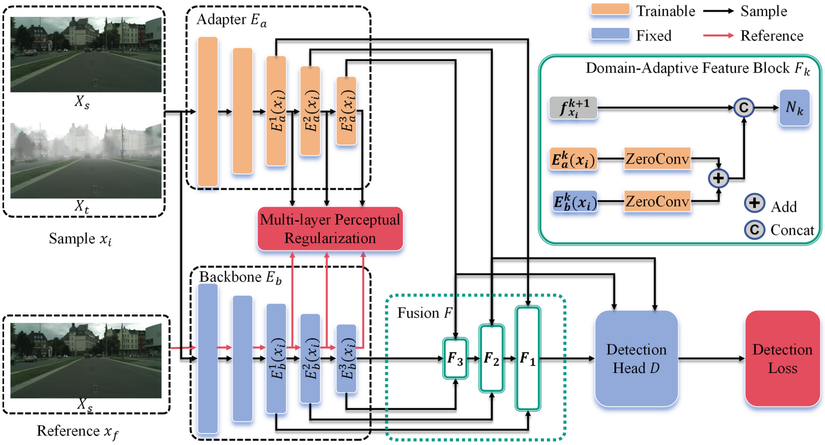


Fig. 2. The framework of proposed EIF-YOLO. The adapter E_a is initialized with pre-trained YOLOv8 parameters and extracts features from different domain data. We introduce multi-layer perceptual regularization to ensure that the extracted features from the adapter more focus on the domain-invariant information. Additionally, we enabled a domain-adaptive feature fusion, for improving the model's performance across different environments.

2 Related Work

2.1 Object Detection

Target detection has witnessed remarkable advancements in recent years, fueled by deep learning and the proliferation of large-scale datasets. Convolutional Neural Networks (CNNs), particularly advanced architectures like YOLO [1], SSD [3], and Faster R-CNN [2], have revolutionized the field by achieving unprecedented accuracy and efficiency. As research progresses, techniques such as attention mechanisms detector DETR [2], multi-scale feature fusion [32,33], and anchor-free detector YOLOX [34] approaches are being explored to enhance detection performance further.

2.2 Domain Adaptation

Domain adaptation, a pivotal concept in machine learning and computer vision, addresses the challenge of deploying models trained on a source domain to operate in a target domain with distinct statistical properties effectively. This paradigm is crucial for bridging the gap between synthetic and real-world data, enabling models trained in controlled environments to be applied to complex, real-world scenarios.

Recent work can be divided into two mainstream directions: Unsupervised Domain Adaptation (UDA) and Semi-supervised or Weakly-supervised Domain Adaptation. Typical UDA methods [11] focus on leveraging unlabeled target data to align the distributions between the source and target domains; some studies have proposed UDA methods that combine self-training [12] and pseudo labeling [13] to generate pseudo labels based on the model's predictions. In addition, some studies have explored how to use adversarial training [28] to improve the domain adaptation performance further [29], by introducing a domain discriminator to encourage the model to learn more domain invariant feature representations. Semi-supervised approaches build upon UDA by incorporating limited labeled data from the target domain to guide the adaptation process further [14,15].

2.3 Domain Adaptation Object Detection

Recent works on domain adaptation have focused on various approaches to bridge the gap between the source and target domains for object detection. The main idea is to leverage unlabeled or weakly labeled data from the target domain to align the feature distributions between the two domains.

A typical method is to embed an image enhancement module to align the feature distributions firstly [16–18,20]. The design of these image enhancement modules usually follows the atmospheric scattering model [22], and is trained by detection loss. However, additional modules increase the model's size while slowing down the inference speed [19,21] achieved faster inference speed by optimizing the network structure. [40] combining self-training with image training modules to improve the model's accuracy.

Another method explores the use of adversarial learning to achieve domain adaptation. This approach effectively reduces the domain shift for object detection tasks [23,24,26]. After the gradient reversal layer (GRL) [28] is discovered, it can be used to assist domain discriminator align features in an adversarial way [27,29,30]. The popular student-teacher framework is used to pseudo-label target domain images; providing pseudo-labels in the target domain will improve the detection capability of the detector in the target domain [27,29,30].

3 Methodology

3.1 Preliminaries

Our model builds upon YOLOv8, a widely used single-stage detection model. The detection model consists of three main components: backbone, multiple neck blocks, and multiple detection heads.

The backbone of the model extracts features at multiple levels, including low, medium, and high levels. To handle those features, YOLOv8 incorporates multiple neck blocks that employ up-sampling and down-sampling strategies to fuse features from different levels. The multiple detection head decodes the fused features to obtain object localization and recognition results, including the coordinates of bounding boxes and corresponding object categories. Adopting a multi-scale approach can significantly improve the model's capability to detect objects of varying sizes. The detection model is primarily optimized using the following loss:

$$\mathcal{L}_{Det} = L_{coord} + L_{obj} + L_{cbs}, \tag{1}$$

where L_{coord} is a coordinate loss used to determine the position of an object. L_{obj} is the object loss, also known as the confidence loss, used to calculate the confidence of the prediction box. L_{cbs} is a category loss used to calculate the category of the prediction box.

3.2 Overview

Given the source domain datasetX_s, we refer to [17,25] add environmental perturbations to X_s to obtain the paired target domain dataset X_t. To ensure that the model maintains its performance on the source domain, we create a training dataset $X = \{x_i \mid x_i \in X_s \cup x_i \in X_t\}$ by combining data from source domain and target domain in a ratio of 1:3.

In our work, we aim to learn more domain invariant information while preserving the model's excellent performance in source domain feature extraction. Firstly, to provide the model with additional capacity to learn domain invariant information, we reused the original backbone E_b as an environment-independent adapter, noted as E_a. The multi-layer features extracted from x_i using E_a and E_b are represented as $\{E_a^k(x_i)\}_{k=1}^K$ and $\{E_b^k(x_i)\}_{k=1}^K$ respectively, where K is 3 and represents the features of the low, medium and high level in order from smallest to largest. Meanwhile, in order to constrain E_a to focus on domain invariant information, we additionally input a reference image x_f, which corresponds to x_i, into E_b to obtain multi-layer feature maps, and we deployed a multi-layer perceptual regularization to regularization on E_b by optimizing the distance between features extracted by different feature extractors. Subsequently, We converted the original neck module into a domain-adaptive fusion module $F = \{F_k\}_{k=1}^{K}$, which can provide dynamic weights for feature fusion and get fused multi-layer feature maps $\{f_{x_i}^{k}\}_{k=1}^{K}$. Finally, the detection head D decodes fused features and obtains the object localization and recognition results. During training, the original pretrained backbone E_b, the neck module $N = \{N_k\}_{k=1}^{K}$ embedded in F, and the head D remain frozen.

3.3 Environment-Independent Feature Extraction

In clear images, a pre-trained object detector can accurately identify objects due to the feature extractor's ability to capture object-related information. However, the feature extractor may need help when images are affected by complex environments. To address this issue, we introduce an environment-independent adapter E_a specifically for handling different domain images. This adapter allows it to disregard environmental interference and concentrate on the object of interest. However, training a new feature extractor from scratch is a laborious task and can introduce conflicts with the features extracted by the original feature extractor E_b, posing challenges for subsequent fusion modules. Hence, we proposed to reuse the pre-trained feature extractor as an adapter by copying its structure and weights. We optimize E_a by inputting x_i, while the input of E_b is x_f, and the E_b is frozen. This approach enables the E_a to learn the object-related features of abnormal images while leveraging the knowledge acquired by the E_b.

While the detection loss of training data can guide the E_a to extract object features in input images, it overlooks the invariant information shared between different domains, potentially leading to model overfitting. To address this, we introduce the multi-layer perceptual regularization that guides E_a to reduce the interference of complex backgrounds on objects at different layers. We leverage the superior detection performance of E_b on clear reference images and the consistency of labels. Specifically, we added a reference image x_f for E_b to extract feature maps. By employing mean squared error (MSE) loss between $\{E_a^k(x_i)\}_{k=1}^{K}$ and $\{E_b^k(x_f)\}_{k=1}^{K}$, E_a will be constrained to reduce the gap between them, which guide E_a to extract features that primarily focus on the objects. The multi-layer perceptual regularization is defined as follows:

$$\mathcal{L}_{MPR} = \sum_{k=1}^{K} MSE(E_a^k(x_i), E_b^k(x_f)), \tag{2}$$

The value of K represents the number of multi-scale, and by default, it remains the same as YOLOv8 and is set to 3. By applying the aforementioned constraints, we can utilize the object-related features extracted from E_b to guide E_a in disregarding the environmental influence.

3.4 Domain-Adaptive Feature Fusion

The features obtained by different extractors may contain complementary information. However, when applying the pre-trained frozen feature extractor E_b to an abnormal image, its performance relies solely on the extent of interference caused by environmental factors. Directly fusing the different features could compromise the model's robustness. Zero-Initialized Layers, a convolutional layer with initial weights and bias values set to zero, are used in ControlNet [31] for connecting different features. We transform the original Neck module into domain-adaptive fusion F to merge features from different extractors efficiently, the detailed structural design of DAF can refer to the Sect. 3 of Supplementary Materials. This module learns fusion weights to combine features from different feature extractors dynamically:

$$f_{x_i}^{k} = F_k(E_a^k(x_i), E_b^k(x_i), f_{x_i}^{k+1}), k = 1, 2, 3, \tag{3}$$

where $f_{x_i}^{k+1}$ denotes the original feature, when $k = 3$, the original feature is equal to $E_b^3(x_i)$. To achieve the adaptive fusion of features, those extracted by different feature extractors are combined with those dynamic weights and then added to the original data stream as residuals. This process enables improved model performance on the target domain while preserving the detection performance on the source domain data to the greatest extent possible.

3.5 Training and Inference

During training, we input the same image x_i into two feature extractors: adapter E_a and original backbone E_b. Additionally, we input the corresponding clear reference image x_r into the E_b to apply MPR. The final loss, computed against the target domain data, is utilized to optimize the adapter E_a and the fusion module F. The optimization formula for the entire model is as follows:

$$\begin{aligned} &\min_{E_a} \mathcal{L}_{Det} + \omega \mathcal{L}_{MPR}, \\ &\min_{F} \mathcal{L}_{Det}, \end{aligned} \tag{4}$$

where ω represents the coefficient of the multi-layer perceptual regularization. After completing the training process, the object detection model can utilize data from the source or target domain as input for object detection. The pseudocode for the training process can be referred to in Sect. 4 of the Supplementary Materials.

4 Experiments

The performance of EIF-YOLO is evaluated through experiments conducted on two common tasks: foggy detection and low-light detection. We begin by presenting comprehensive information regarding the benchmarks and experimental setups. Next, we quantitatively compare EIF-YOLO with several state-of-the-art methods to assess its performance. Additionally, we provide visual comparisons to demonstrate the direct effectiveness of EIF-YOLO in complex environments. Moreover, we perform ablation experiments to validate the efficacy of each module in the proposed model.

4.1 Experimental Settings

Datasets. The source domain clear images are derived from the VOC [35] dataset (VOC2007 and VOC2012), which consists of 21,503 traffic scene images with accurately labeled objects belonging to 20 object classes for the object detection task. In the foggy detection task, we evaluate the model's robustness using the real-world hazy dataset called RTTS [38]. This dataset comprises 4,322 natural hazy images with five labeled object classes: person, bicycle, car, bus, and motorcycle. To ensure consistency, we select VOC images containing these five categories and create a subset called VOC_nf. Following a method described in [17], we use the atmospheric scattering formula to generate images in foggy conditions, and use power operations to generate images in low light conditions. We convert VOC_nf to the target domain dataset VOC_f, which contains images captured in foggy conditions with varying degrees of occlusion. For the low-light detection task, we introduce the part of ExDark [37] dataset as the val set, which consists of 2,546 images containing 10 object categories, including bike, boat, bottle, bus, car, cat, chair, dog, motorbike, and person. To train the low-light domain adaptation object detection, we filter the VOC dataset and obtain VOC_nd. Subsequently, we convert VOC_nd into low-light datasets VOC_d using a reference method [17]. The composition details of the datasets are provided in Table 1.

Hyperparameter. To ensure training stability, we adopt hyperparameter settings based on the default configurations of YOLOv8. We initialize EIF-YOLO using the pre-trained model on the COCO [36] dataset. The optimization process utilizes the AdamW optimizer with a learning rate of 0.000119 and a momentum of 0.9 The value of ω in Eq. (4) is set to 10. The model is trained for 100 epochs with a batch size of 64. In order to maintain the model's performance on the source domain, the input data consists of a mixture of source and target domain data in a 1:3 ratio, the experimental results on mixed data can refer to Sect. 1 of Supplementary Materials. All experiments are conducted on an NVIDIA TITAN X GPU.

Evaluation Metrics. The mAP (%), standing for Mean Average Precision, is a crucial metric used to evaluate the performance of object detection models. It averages the AP (Average Precision) scores across all object categories.

Table 1. The normal datasets, VOC_nf and VOC_nd, consist of the same data as their corresponding generated datasets VOC_f and VOC_d, respectively.

Dataset	Images		Instances										
	train	val	person	bicycle	car	motor	bus	boat	cat	dog	bottle	chair	Total
RTTS	–	4322	7950	534	18413	1838	862	–	–	–	–	–	29577
VOC_f	8111	1712	3286	244	921	235	151	–	–	–	–	–	4852
ExDark	–	2546	2802	464	1105	461	279	564	366	402	573	643	7665
VOC_d	12509	3831	5227	389	1541	369	254	393	370	530	657	1374	11104

4.2 Comparative Experiments on Foggy Images

To evaluate the effectiveness of EIF-YOLO in foggy detection, we conduct a comparative analysis involving baseline YOLOv8 and other methods. IA-YOLO [17], GDIP-YOLO [19], DENet [39] and TIENet [41] trained an image enhancement module. BAD-Net [40], HLA-HOD [47], and HLNet [46] trained additional modules for aligning feature maps. SDA [45] is an unsupervised domain adaptation method. To ensure a fair comparison, we maintained consistency by using the same training sets for all the methods during the training process.

Table 2 compares mAP between EIF-YOLO and other methods on the three test sets. The baseline method performs well on the source domain data but needs help to achieve satisfactory results on the untrained target domain data. Specifically, it only achieves 46.2% mAP on the foggy weather dataset VOC_f, indicating significant performance degradation in foggy conditions. The other comparison methods show improved performance in detecting objects under foggy conditions by training on the target domain. However, their performance in the source domain is adversely affected to varying degrees. In contrast, our model exhibits an improvement of 1.58% points. in detection performance on the VOC_nf compared to the baseline. This improvement is attributed to the introduction of domain-adaptive feature fusion and the training of the adapter with a mixture of source and target domain images. These adaptations enable the model to fuse features adapted by different feature extractors effectively. Furthermore, by optimizing the adapter, our model demonstrates improvements of 38.58 and 6.3% points on the generated dataset VOC_f and the real-world hazy dataset RTTS, respectively, compared to the sub-optimal method. This result demonstrates the efficacy of our model in addressing the DAOD task for foggy scenes.

Additionally, we present a visual comparison between EIF-YOLO and the baseline (YOLOv8) in Fig. 3. The adapter in EIF-YOLO focuses on extracting environment-independent object features through the optimization of detection loss and multi-layer perceptual regularization. For more representative results, please refer to Sect. 2 of the Supplementary Materials.

Table 2. Performance comparison on foggy images in terms of mAP.

Method	VOC_nf_val	VOC_f_val	RTTS
Baseline	86.40	46.20	51.30
IA-YOLO [17]	73.23	72.03	37.80
GDIP-YOLO [19]	75.36	73.37	42.84
BAD-Net [40]	85.86	85.58	53.15
DENet [39]	84.13	83.56	53.70
TIENet [41]	–	77.50	52.10
HLA-HOD [47]	–	83.43	56.90
SDA [45]	–	–	35.10
HLNet [46]	–	–	55.60
EIF-YOLO	**87.98**	**85.78**	**57.60**

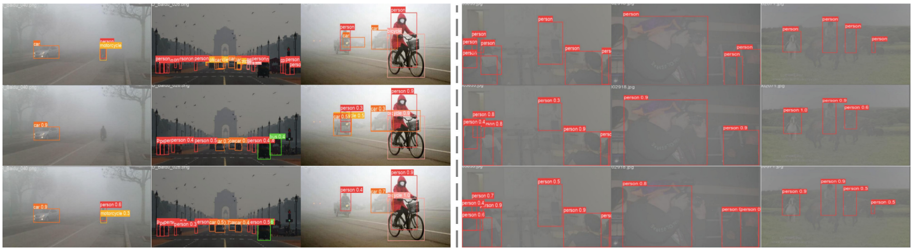


Fig. 3. Representative results of EIF-YOLO(row3) and the baseline(row2) on VOC_f_val (Columns 1 to 3) and RTTS (Columns 4 to 6). The first row is the ground truth.

4.3 Comparative Experiments on Low-Light Images

In the low-light detection scenario, we conducted a similar comparison between our model and other comparison methods. LAR-YOLO [42] trained an image enhancing module, and SC-Det [48] trained an additional module for aligning feature maps. LIDA-YOLO [49] and T2 [43] are unsupervised domain adaptation methods. The remaining experimental settings remained consistent with those used for foggy detection.

Table 3 presents the mAP results for each method. Notably, our model demonstrates improved performance even on the derived data, highlighting its effectiveness. A significant improvement is observed on both low-light datasets, especially on VOC_d, where we achieve a 9.2% points improvement over the baseline. This result demonstrates the efficacy of our method in low-light detection tasks.

Furthermore, we provide a visual comparison with the baseline(YOLOv8) in Fig. 4. Our model accurately recognizes objects even in extremely poor lighting conditions, emphasizing its ability to capture environment-independent features for detecting objects of interest. For more representative results, please refer to Sect. 2 of the Supplementary Materials.

Table 3. Performance comparison on low-light images in terms of mAP.

Method	VOC_nd_val	VOC_d_val	ExDark
Baseline	79.10	66.40	62.10
IA-YOLO [17]	70.20	59.40	40.37
GDIP-YOLO [19]	63.23	57.85	42.56
LAR-YOLO [42]	74.49	62.79	42.58
DENet [39]	73.17	67.81	51.51
SC-Det [48]	69.00	–	63.00
LIDA-YOLO [49]	–	–	56.65
T2 [43]	–	–	61.76
EIF-YOLO	**79.20**	**75.66**	**65.70**

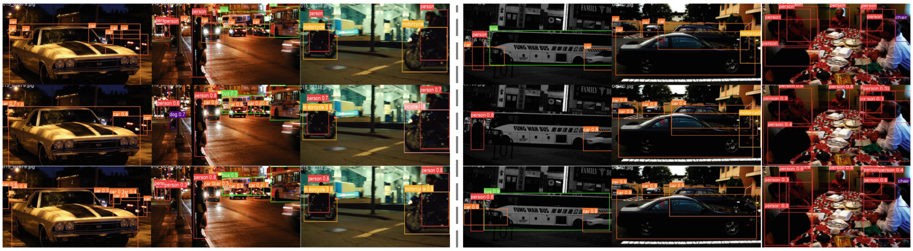


Fig. 4. Representative results of EIF-YOLO and the baseline (YOLOv8) on VOC_d_val (Columns 1 to 3) and ExDark (Columns 4 to 6). The first row is ground truth, the second row is the baseline method (YOLO v8), and the third row is EIF-YOLO.

4.4 Ablation Study

To assess the effectiveness of each module in our proposed method, we conducted ablation experiments under various settings, including environment-independent adapter, multi-layer perceptual regularization, and domain-adaptive fusion. These experiments were performed on three datasets. The results are presented in Table 4, with the first row indicating the baseline (YOLOv8) performance.

For the foggy detection, it is evident that the introduction of the adapter and MPR improves the detection performance on VOC_f by about 37.94% points. However, it leads to a degradation in the detector's performance on VOC_nf and RTTS, especially on VOC_nf, decline 42.86% points. This result is based on the fact that RTTS experiences fewer disturbances from foggy weather compared to VOC_f. It primarily exhibits a domain gap with the source domain data. Consequently, fine-tuning the model on VOC_f could potentially undermine the detector's original capability. Furthermore, simple feature fusion disregards conflicts between different features, resulting in performance degradation. By incorporating domain-adaptive fusion, the model's performance demonstrates significant improvement across all three test sets, particularly on the source domain data, with a noteworthy increase of 44.44% points (Table 4)

Table 4. Ablation analysis on VOC_nf_val, VOC_f_val and RTTS in terms of mAP.

Method	EIA	MPR	DAF	VOC_nf_val	VOC_f_val	RTTS
Baseline				86.40	46.20	51.30
EIF-YOLO	√	√		43.54	84.14	47.60
	√		√	71.40	81.40	53.60
	√	√	√	**87.98**	**85.78**	**57.60**

4.5 Impact of Hyper-parameters

The coefficient ω in Eq. (4), which corresponds to mutil-layer regularization, plays a crucial role in the training process. We experiment with the performance of the model in various datasets when ω is set to different values of {0, 0.5, 1, 2, 5, 7.5, 10, 15}. In Fig. 5, it is observed that the results are similar when the ω is small compared to when MPR is not included. However, as the MPR value increases, the overall performance of the model gradually improves. When it reaches 15, the model performance starts to decline. Therefore, we use $\omega = 10$ to achieve the optimal performance of the model.

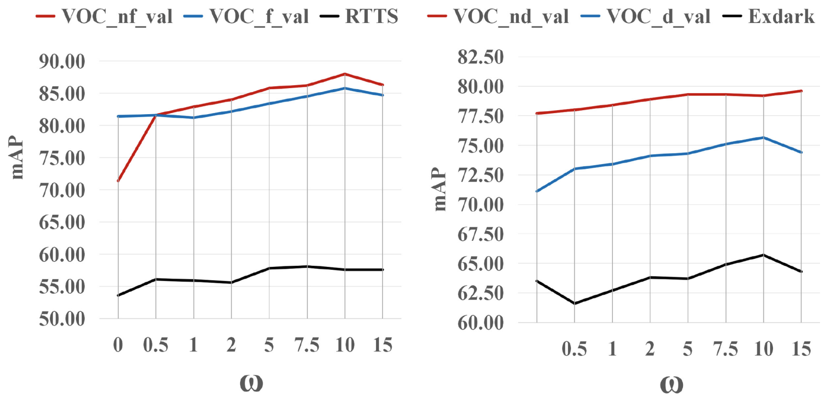


Fig. 5. The impact of coefficient ω on foggy detection (left) and low-light detection (right) in terms of mAP.

5 Conclusion

We propose EIF-YOLO, a simple yet efficient method that enhances robust object detection in adverse environments by fusing environment-independent features. We ultimately retain the parameters of pretrained YOLOv8 and get a learnable adapter responsible for detecting abnormal images by copying the original backbone of YOLOv8. To enable the adapter to focus on object-specific features, we introduce multi-layer perceptual regularization at multiple scales. This guidance encourages that the features extracted by the adapter align closely with those obtained from clear reference images. Directly tuning the original object detector using the target domain images adversely impacts the performance in the source domain. To address these issues, we introduce domain-adaptive fusion, which enables the model to preserve performance in the source domain while simultaneously improving results in the target domain. Experimental results demonstrate that our method reduces the influence of complex environments on the detection performance.

Acknowledgements. This work was supported in part by the National Natural Science Foun dation of China (Project No. 62072189), in part by the Guang-Dong Basic and Applied Basic Research Foundation (Project No. 2022A1515011160, 2024A1515011437), and in part by TCL Science and Technology Innovation Fund (Project No. 20231752).

References

1. Redmon, J., Divvala, S., Girshick, R., Farhadi, A.: You only look once: unified, real-time object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 779–788 (2016)

2. Ren, S., He, K., Girshick, R., Sun, J.: Faster r-cnn: towards real-time object detection with region proposal networks. In: Advances in Neural Information Processing Systems, vol. 28 (2015)
3. Liu, W., et al.: SSD: single shot multibox detector. In: Proceedings of the European Conference on Computer Vision, Amsterdam, The Netherlands, 11–14 October 2016 (2016)
4. Li, W., Li, F., Luo, Y., et al.: Deep domain adaptive object detection: a survey. In: 2020 IEEE Symposium Series on Computational Intelligence (SSCI), pp. 1808–1813. IEEE (2020)
5. Gopalan, R., Li, R., Chellappa, R.: Domain adaptation for object recognition: an unsupervised approach. In: 2011 International Conference on Computer Vision, pp. 999–1006. IEEE (2011)
6. Ben-David, S., Blitzer, J., Crammer, K., et al.: A theory of learning from different domains. Mach. Learn. **79**, 151–175 (2010)
7. Vaswani, A., Shazeer, N., Parmar, N., et al.: Attention is all you need. In: Advances in Neural Information Processing Systems, vol. 30 (2017)
8. Rabbi, J., Ray, N., Schubert, M., et al.: Small-object detection in remote sensing images with end-to-end edge-enhanced GAN and object detector network. Remote Sens. **12**(9), 1432 (2020)
9. Li, Y., Hou, Q., Zheng, Z., et al.: Large selective kernel network for remote sensing object detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 16794–16805 (2023)
10. He, C., Li, K., Zhang, Y., et al.: Camouflaged object detection with feature decomposition and edge reconstruction. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 22046–22055 (2023)
11. Xie, Q., Dai, Z., Hovy, E., et al.: Unsupervised data augmentation for consistency training. In: Advances on Neural Information Processing System, vol. 33, pp. 6256–6268 (2020)
12. Liu, H., Wang, J., Long, M.: Cycle self-training for domain adaptation. In: Advances on Neural Information Processing System, vol. 34, pp. 22968–22981 (2021)
13. Li, J., Zhou, K., Qian, S., et al.: Feature re-representation and reliable pseudo label retraining for cross-domain semantic segmentation. IEEE Trans. Pattern Anal. Mach. Intell. **46**(3), 1682–1694 (2022)
14. Vale, K.M.O., Gorgônio, A.C., Flavius Da Luz, E.G., et al.: An efficient approach to select instances in self-training and co-training semi-supervised methods. IEEE Access **10**, 7254–7276 (2021)
15. Lu, X., Wu, J., Huang, J., et al.: Co-training-teaching: a robust semi-supervised framework for review-aware rating regression. ACM Trans. Knowl. Discov. Data **18**(2), 1–16 (2023)
16. Qiu, Y., Lu, Y., Wang, Y., et al.: IDOD-YOLOV7: image-dehazing YOLOV7 for object detection in low-light foggy traffic environments. Sensors **23**(3), 1347 (2023)
17. Liu, W., Ren, G., Yu, R., et al.: Image-adaptive YOLO for object detection in adverse weather conditions. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 36, no. 2, pp. 1792–1800 (2022)
18. Liu, T., Zhang, Z., Lei, Z., et al.: An approach to ship target detection based on combined optimization model of dehazing and detection. Eng. Appl. Artif. Intell. **127**, 107332 (2024)
19. Kalwar, S., Patel, D., Aanegola, A., et al.: Gdip: gated differentiable image processing for object detection in adverse conditions. In: 2023 IEEE International Conference on Robotics and Automation (ICRA), pp. 7083–7089. IEEE (2023)

20. Tzeng, E., Hoffman, J., Saenko, K., et al.: Adversarial discriminative domain adaptation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 7167–7176 (2017)
21. Hu, K., Wu, F., Zhan, Z., et al.: High-low level task combination for object detection in foggy weather conditions. J. Vis. Commun. Image Represent. **98**, 104042 (2024)
22. Narasimhan, S.G., Nayar, S.K.: Vision and the atmosphere. Int. J. Comput. Vision **48**, 233–254 (2002)
23. Chen, Y., Li, W., Sakaridis, C., et al.: Domain adaptive faster r-cnn for object detection in the wild. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3339–3348 (2018)
24. Huang, W.J., Lu, Y.L., Lin, S.Y., Xie, Y., Lin, Y.Y.: Aqt: adversarial query transformers for domain adaptive object detection. IJCAI-ECAI (2022)
25. Li, J., Xu, R., Ma, J., et al.: Domain adaptive object detection for autonomous driving under foggy weather. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pp. 612–622 (2023)
26. Guan, D., Huang, J., Xiao, A., et al.: Uncertainty-aware unsupervised domain adaptation in object detection. IEEE Trans. Multimedia **24**, 2502–2514 (2021)
27. Li, Y.J., Dai, X., Ma, C.Y., et al.: Cross-domain adaptive teacher for object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 7581–7590 (2022)
28. Ganin, Y., Lempitsky, V.: Unsupervised domain adaptation by backpropagation. In: International Conference on Machine Learning, pp. 1180–1189. PMLR (2015)
29. Cai, Q., Pan, Y., Ngo, C.W., et al.: Exploring object relation in mean teacher for cross-domain detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 11457–11466 (2019)
30. Deng, J., Li, W., Chen, Y., et al.: Unbiased mean teacher for cross-domain object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 4091–4101 (2021)
31. Zhang, L., Rao, A., Agrawala, M.: Adding conditional control to text-to-image diffusion models. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 3836–3847 (2023)
32. Wang, W., Xie, E., Song, X., et al.: Efficient and accurate arbitrary-shaped text detection with pixel aggregation network. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 8440–8449 (2019)
33. Lin, T.Y., Dollár, P., Girshick, R., et al.: Feature pyramid networks for object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2117–2125 (2017)
34. Ge, Z., Liu, S., Wang, F., et al.: Yolox: exceeding yolo series in 2021. arXiv preprint arXiv:2107.08430 (2021)
35. Everingham, M., Van Gool, L., Williams, C.K., Winn, J., Zisserman, A.: The pascal visual object classes (voc) challenge. Int. J. Comput. Vision **88**(2), 303–338 (2010)
36. Lin, T.-Y., et al.: Microsoft COCO: common objects in context. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) ECCV 2014. LNCS, vol. 8693, pp. 740–755. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-10602-1_48
37. Loh, Y.P., Chan, C.S.: Getting to know low-light mages with the exclusively dark dataset. Comput. Vision Image Understand. **178**, 30–42 (2019)
38. Li, B., et al.: Benchmarking single-image dehazing and beyond. IEEE Trans. Image Process. **28**(1), 492–505 (2018)

39. Qin, Q., Chang, K., Huang, M., et al.: DENet: detection-driven enhancement network for object detection under adverse weather conditions. In: Proceedings of the Asian Conference on Computer Vision, pp. 2813–2829 (2022)
40. Li, C., et al.: Detection-friendly dehazing: object detection in real-world hazy scenes. IEEE Trans. Pattern Anal. Mach. Intell. **45**(7), 8284–8295 (2023). https://doi.org/10.1109/TPAMI.2023.3234976
41. Wang, Y., Guo, J., Wang, R., et al.: TIENet: task-oriented image enhancement network for degraded object detectio. Signal Image Video Process. **18**(1), 1–8 (2024)
42. Yao, M., Lu, Y., Mou, J., et al.: End-to-end adaptive object detection with learnable Retinex for low-light city environment. Nondestruct. Test. Evaluat. **39**(1), 142–163 (2024)
43. Cui, X., Ma, L., Ma, T., et al.: Trash to treasure: low-light object detection via decomposition-and-aggregation. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 38, no. 2, pp. 1417–1425 (2024)
44. Ye, J., Wu, Y., Peng, D.: Low-quality image object detection based on reinforcement learning adaptive enhancement. Pattern Recogn. Lett. **182**, 67–75 (2024)
45. Zhou, Q., Gu, Q., Pang, J., et al.: Self-adversarial disentangling for specific domain adaptation. IEEE Trans. Pattern Anal. Mach. Intell. **45**(7), 8954–8968 (2023)
46. Hu, K., Wu, F., Zhan, Z., Luo, J., Pu, H.: High-low level task combination for object detection in foggy weather conditions. J. Visual Commun. Image Represent. **98**, 104042 (2024)
47. Shen, Y., Yu, R., Shu, N., et al.: HLA-HOD: joint high-low adaptation for object detection in hazy weather conditions. Int. J. Intell. Syst. **2023**(1), 3691730 (2023)
48. Lin, T., Huang, G., Yuan, X., Zhong, G., Huang, X., Pun, C.M.: SCDet: decoupling discriminative representation for dark object detection via supervised contrastive learning. Visual Comput. **40**(5), 3357–3369 (2024)
49. Xiao, Y., Liao, H.: LIDA-YOLO: an unsupervised low-illumination object detection based on domain adaptation. IET Image Proc. **18**(5), 1178–1188 (2024)

Transformer-Based RGB and LiDAR Fusion for Enhanced Object Detection

Reza Sadeghian(✉), Niloofar Hooshyaripour, and WonSook Lee

University of Ottawa, Ottawa, ON, Canada
{r.sadeghian,nhoos082,wslee}@uottawa.ca

Abstract. The integration of LiDAR and camera data has demonstrated significant potential in enhancing the accuracy and robustness of object detection systems. Therefore, developing a proficient fusion technique for these modalities is vital to harness their combined strengths. In this study, we propose TransfuseNet, a novel Transformer-based method for 2D object detection that deviates from the usual use of transformers in cross-attention tasks. This approach emphasizes self-attention to efficiently integrate camera and LiDAR inputs, thereby enhancing global context synthesis from both sources. Our designed Transformer architecture processes multi-modal feature maps derived from LiDAR and image data, which improves feature extraction and contextual understanding. Additionally, we examined different fusion operators, focusing on their roles in the later stages of fusion. This analysis led to the creation of Multi-Convolutional Fusion (MCF), a new strategy that uses a priority gate to highlight features with higher importance scores during fusion. Experimental results on KITTI benchmark datasets demonstrate that our approach not only matches state-of-the-art methods but is also significantly faster, making it ideal for rapid decision-making in autonomous driving.

Keywords: Object Detection · Multi-modal Fusion · Transformer

1 Introduction

The advancement of autonomous driving technology heavily relies on accurate and robust object detection methods, enabling vehicles to perceive and comprehend their surroundings. Image-based object detection methods either one-stage detector [10,22] or two-stage detectors [8,16,28] often struggle under challenging conditions such as adverse lighting and occlusions, significantly impairing visual clarity and detection accuracy (Fig. 1).

Integrating LiDAR data has emerged as a promising solution to these limitations. LiDAR sensors provide precise and dense 3D point cloud information, which complements the visual information captured by cameras, thereby enhancing detection accuracy and robustness in autonomous systems.

However, LiDAR systems are not without their limitations. Environmental factors such as fog or heavy rain can impair LiDAR data, leading to incomplete

A. Antonacopoulos et al. (Eds.): ICPR 2024, LNCS 15329, pp. 445–460, 2025.
https://doi.org/10.1007/978-3-031-78110-0_29

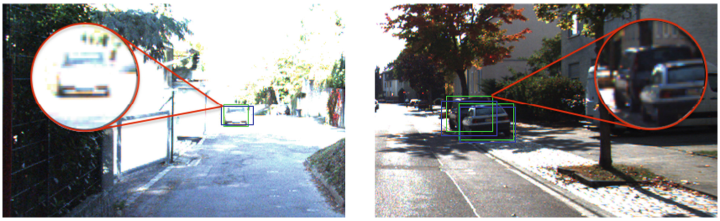


Fig. 1. Left: Poor Illumination. Right: High Occlusion. Both conditions illustrate object detection challenges effectively addressed by our proposed network. Bounding boxes in green and blue denote predictions and ground truth, respectively. (Color figure online)

or inaccurate 3D detections. This presents a critical challenge for autonomous driving, where accurate detection is vital. In critical scenarios like collision avoidance, the need for accurate and rapid detection surpasses the dimensionality of the detection, emphasizing the importance of lightweight and fast networks.

To address these needs, we propose TransfuseNet, a network designed to enhance safety by swiftly identifying potential hazards. Our focus on 2D object detection within TransfuseNet strategically balances computational efficiency with detection accuracy, meeting the real-time demands of autonomous driving where rapid response and decision-making are essential.

Several fusion approaches have been proposed to integrate LiDAR and image data for object detection. One common approach is fusing the two modalities at the feature level through concatenation or addition. However, these methods often fail to capture the complex spatial dependencies present in LiDAR data, limiting their effectiveness in certain scenarios.

Motivated by the potential of Transformers, we introduce a novel Transformer-based network tailored for fusing LiDAR and image data, a pioneering effort in the 2D object detection task. Previous methodologies [1,17] have primarily engineered Transformers for 3D detection, often relegating 2D detection to a secondary outcome. Recognizing the importance of targeting 2D detection, our method underscores its strategic significance in achieving a fine balance between computational efficiency and detection accuracy which is vital for addressing the real-time requirements of autonomous driving. Moreover, contrary to these methods that primarily applied Transformers to generate proposals or utilized cross-attention techniques, our approach distinctively leverages the self-attention mechanism of Transformers to integrate camera and LiDAR data, enabling a comprehensive synthesis of global context from both modalities.

In constructing a robust sensor fusion framework, the optimal selection of a fusion operator, especially for late fusion, is essential to integrate diverse modalities and their associated feature maps effectively. The recent literature has proposed several fusion operators, ranging from simple element-wise, non-learnable operators to more advanced learnable approaches [35,36] that capture the inherent relationships between features. However, The effectiveness of intricate fusion operators must be assessed before introducing computational complexity into a network.

To evaluate the effectiveness of our proposed approach, we conduct extensive experiments on KITTI benchmark datasets [9] for 2D object detection in autonomous driving scenarios.

Our primary contributions can be summarized as follows:

- We introduce TransfuseNet, which employs Transformer self-attention for fusing camera and LiDAR data, demonstrating state-of-the-art results in 2D object detection.
- TransfuseNet is designed to be simple and fast, making it suitable for integration within autonomous driving systems, particularly for edge computing applications.
- We present a novel learnable fusion method called MCF, that leverages priority gates, demonstrating superior performance over the fusion techniques evaluated in this study.
- The effectiveness of various fusion operators for 2D object detection is thoroughly investigated. Our systematic exploration highlights the impact of different fusion strategies on the overall detection performance, providing valuable insights for future research in this domain.

The remainder of this paper is organized as follows: Sect. 2 provides an overview of related work in object detection using data fusion. Section 3 describes our proposed Transformer-based fusion approach and the different fusion operators employed as late fusion in this study. The experimental results are presented in Sect. 4, followed by the conclusion and future research directions in Sect. 5.

2 Related Work

Camera-only Detectors. In autonomous driving, there has been a substantial focus on detecting objects using only camera inputs, a trend largely driven by the KITTI benchmark. Due to the KITTI dataset's reliance on a singular front camera, this has prompted the development of various methods specifically tailored for monocular 3D detection, as referenced in several studies [2,19,30].

LiDAR-only Detectors. Initial strategies for object detection via point clouds can be divided into two primary categories. One approach simplifies the point cloud into more streamlined forms, such as Bird's Eye View (BEV) images [24,25], Frontal View (FV) images [3,7], and three-dimensional attribute representations [39]. Another approach involves utilizing the raw point clouds directly [12,31,32]. In this context, [38] proposes an efficient single-stage point-based 3D detector, focusing on foreground points as key for object detection. It employs innovative, instance-aware downsampling strategies to selectively prioritize these foreground points in detecting objects of interest. [32] presents the Point-Voxel Transformer (PVT) module, which combines voxel-based feature encoding with an innovative query initialization module. This approach efficiently fuses long-range voxel contexts and precise point positions, effectively integrating contextual and geometric data. BtcDet [31] effectively estimates occluded object shapes in point clouds. It identifies areas affected by occlusion, predicts occupancy probabilities to detect object shapes, and integrates this data to produce refined 3D proposals and final bounding boxes.

Camera-LiDAR Fusion Detectors. In object detection, combining LiDAR and camera data is becoming increasingly common due to the unique advantages each provides. This approach falls into two categories based on how the data from the two sources is fused. The first category is point-level fusion. Here, features from images are linked with raw LiDAR points, and these combined features are then added back as enhanced point data, as seen in studies [13,29]. The second category is feature-level fusion. In this method, LiDAR points are first converted into a feature format [18,37] or used to create initial detection proposals [1]. Then, camera data is associated with these features or proposals, enhancing the overall feature quality, as shown in [4]. TransFusion [1] employs a Transformer-based detection head. The initial layer of this decoder generates preliminary bounding boxes from a LiDAR point cloud through a limited number of object queries. Subsequently, its second layer of the decoder skillfully merges these object queries with valuable image features, utilizing both spatial and contextual correlations. DeepFusion [17] utilizes cross-attention to dynamically identify the relationships between image and LiDAR features during the fusion process.

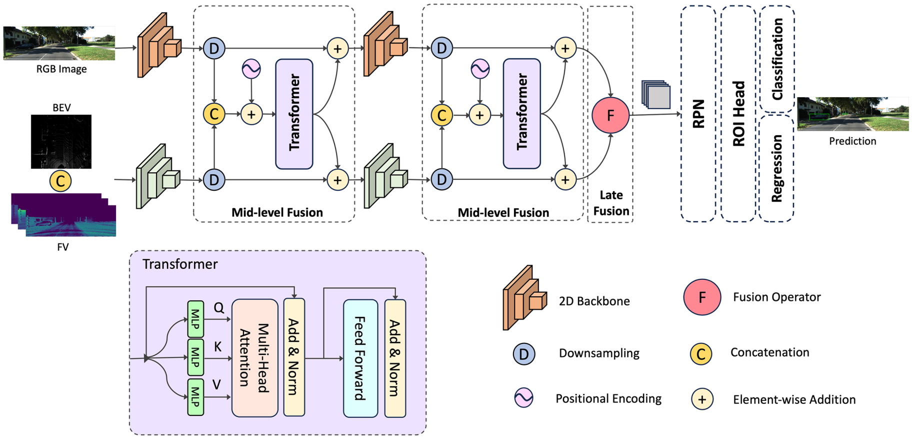


Fig. 2. The overall architecture of our proposed TransfuseNet with single-view RGB and LiDAR BEV/FV inputs. The system employs multiple Transformer layers for intermediate feature map fusion, followed by a late fusion operator. These fused features are input to a Region Proposal Network and a subsequent detection head for bounding box prediction.

3 TransfuseNet

TransfuseNet is a two-stage, end-to-end object detection framework utilizing data fusion. As shown in Fig. 2, it comprises two main parts:(1) A Transformer-based fusion module followed by late fusion operation, integrating data from LiDAR and camera streams for a cohesive representation, and (2) a region proposal and detection head. In the upcoming sections, we will elaborate on the input representation and describe each network component.

3.1 Input Representation

Our research incorporates two distinct sensor types: camera RGB images and LiDAR point cloud data. These sensor inputs are transformed into the FV/BEV format for further processing.

The input for the camera stream is RGB images of KITTI dataset. KITTI dataset provides images with dimensions of 1242×375×3, which will be processed by min-max normalization to scale the values of pixels in the range of 0 to 1. On the other hand, the LiDAR information will be represented in two ways: Bird's Eye View (BEV) and Frontal View (FV). These two technique encodes the 3D point cloud into a 2D image, simplifying object detection.

BEV representation provides a concise yet comprehensive view by indicating the presence or absence of LiDAR points in each corresponding cuboid. Despite reducing its spatial dimensions, it offers a precise depiction of the 3D space. The representation spans a physical space of $[0, 100] \times [-30, 30] \times [-0.6, 3.5]$ meters and is discretized into a 1242×375 image with 7 distinct channels dedicated to height information.

FV representation aligns with the camera's perspective and is constructed using accurate camera calibration parameters from the KITTI dataset. We derive three Frontal View features from the sparse LiDAR point clouds: intensity, depth, and height maps. Intensity maps are based on reflectance values, while depth and height maps utilize ratios relative to their dimensions' maximum values. The three features are then concatenated in the channels' dimensions to create a Frontal View feature image of size 1242×375×3. The encoding of the Frontal View is normalized between 0 and 1.

3.2 Fusion Network

Our model integrates two fusion stages: mid-level, employing the Transformer, and late fusion, utilizing either non-learnable or learnable fusion techniques.

Mid-Level Fusion. Our main concept revolves around leveraging the self-attention mechanism found in Transformers [27] to incorporate global context into both image and LiDAR modalities. We leverage the grid structure feature, inspired by previous work [5,21], to maximize the benefits of Transformers for vision tasks. By integrating this feature, we optimize the performance of Transformer-based architectures in vision-related applications. Incorporating feature maps at different levels, each encompassing distinct information leads to improved comprehension of the input by the network when fusing the features from LiDAR and Images. Consequently, the mentioned features are fused at various levels (Fig. 2) to enhance overall performance and understanding.

Given the intermediate-level feature maps of each stream, which are represented as 3D tensors with dimensions $H \times W \times C$, the individual features from each stream are combined through concatenation, resulting in a tensor with dimensions $(2 \times H \times W) \times C$. Subsequently, a learnable positional embedding is added to these concatenated features (Fig. 2). The purpose of incorporating positional embedding is to enhance the network's understanding of the spatial relationship among input tokens. By considering this positional information, the network gains valuable insights into the spatial context of the input data, contributing to improved performance and interpretation. The produced tensor is used as the input to the Transformer, which produces an output tensor of the same size as the input, as shown in Fig. 2. This output is formed into two feature maps, each with $H \times W \times C$. Then, these feature maps are reintroduced into the distinct modality branches using element-wise summing with the existing feature maps. The ConvMixer [26] feature extractors of the image and LiDAR stream, which operate at various resolutions, are repeatedly subjected to this fusion procedure carried out at a single scale. As High spatial resolution feature map processing is computationally expensive, we use average pooling to downsample higher resolution feature maps from the early encoder blocks to a fixed resolution of H = W = 8 before passing them as inputs to the Transformer. We then use bilinear interpolation to upsample the output to the original resolution before element-wise summing with the existing feature maps.

Utilizing modality-specific feature extractors and following dense feature fusion across diverse resolutions, we obtain a refined feature map of dimensions $16 \times 16 \times 256$. These feature maps, derived from the LiDAR and camera data streams, are then strategically channeled into the advanced late fusion process.

Late Fusion. TransfuseNet adopts a sequential fusion strategy, commencing with mid-level fusion and transitioning to late fusion. Within the late fusion phase, we employed either (1) Non-learnable or (2) Learnable fusion operators, enhancing the experimental process.

Non-learnable Fusion employs no learnable weights, resulting in accelerated processing times. We use two non-learnable fusion operators, elemental addition and multiplication, as shown in Eqs. (1) and (2), respectively, like that of recent research articles [6,20].

$$F_{add} = I_{Camera} \oplus I_{LiDAR} \tag{1}$$

$$F_{mul} = I_{Camera} \otimes I_{LiDAR} \tag{2}$$

Learnable Fusion needs more processing power because they have layers with trainable parameters. As was already indicated, employing learnable combinatorial operators has advantages in terms of their learning potential, particularly

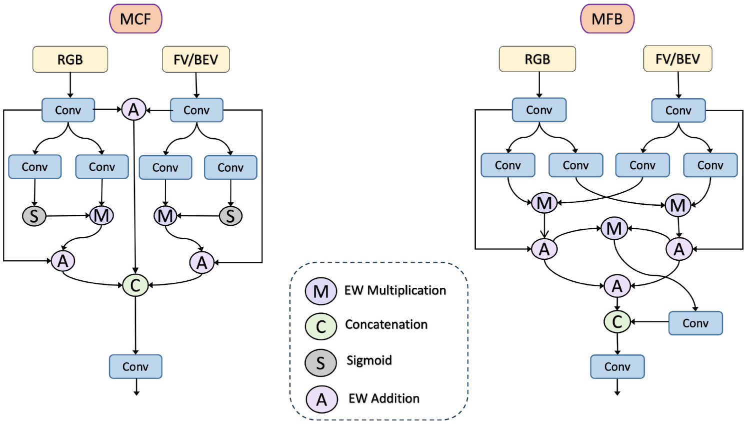


Fig. 3. Learnable fusion operators. Left: Our proposed Multi-Convolutional Fusion operator (MCF). Right: Multi-modal Factorized Bilinear pooling fusion operator (MFB).

in capturing and understanding complex interactions among features generated from various input modalities. As shown in Fig. 3, we also incorporate two learnable fusion operators, Multi-modal Factorized Bilinear pooling (MFB) [36] and a new learnable fusion named Multi-Convolutional Fusion (MCF) that is proposed in this study. MCF leverages the Sigmoid layer as a priority gate, which outputs a probability value for each feature, signifying its importance. Features with higher probability scores are considered more crucial and are thus prioritized during the fusion process with subsequent layers. ReLU is used as the main activation function throughout our model, the padding value is set to 1, and a kernel size of 3 is used for all operations in our network. We use batch normalization techniques to provide stability and efficient training.

3.3 Proposal Generation and Detection Head

Faster R-CNN [22] enhances object detection speed and accuracy by generating region proposals and classifying objects. In our approach, features from the late-fusion output feed into the proposal module, which uses nine anchors based on three scales and aspect ratios. The Region Proposal Network (RPN) predicts objectness scores and bounding box offsets. The detection head module, with three convolution layers and a dropout layer, refines these proposals. Finally, two linear layers classify objects within the refined boxes. During inference, Non-Maximum Suppression (NMS) with an IOU threshold selects high-scoring proposals

4 Experiments

4.1 Dataset and Metric

We evaluate our network using the KITTI object detection benchmark [9], which includes 7,481 training and 7,518 test images. We split the training set into equal training and validation sets, as the test server only assesses 2D detection. The KITTI dataset categorizes labels into three difficulty levels: easy, moderate, and hard. We present the car class accuracy using the Average Precision (AP%) metric with an Intersection-Over-Union (IOU) threshold of 0.7 across all difficulty levels.

4.2 Experimental Setup

We employ the Adam [14] optimizer with a learning rate of 0.001 and a weight decay of 0.00001. The network is trained using a batch size of eight for 150 epochs on an NVIDIA GeForce RTX 3090 GPU. We use ConvMixer [26] model to extract features from the camera RGB images for the image stream. Also, for the LiDAR stream, we use the same model for feature extraction.

4.3 Loss Function

Similar to Faster R-CNN [22], our loss function comprises two integral components. The initial element pertains to the classification loss, for which we employed the cross-entropy loss function to classify the car object. Subsequently, the second component is the regression loss, which exclusively operates when an object is detected. We have adopted the smooth L1 loss function [10] for this purpose.

4.4 Results

As Table 1 shows, we evaluated the performance using various combinations of input modalities and late fusion operators, which will be analyzed in the following. The baseline uses solely RGB data, excluding LiDAR. Through extensive testing, LiDAR's inclusion consistently boosts accuracy across categories, underlining its pivotal role in 2D object detection.

Different Input Modalities. By focusing on a single late fusion operator and comparing the impact of various input modalities, we discover that the choice between BEV and FV representations impacts detection differently. For easily detectable objects (falling in the easy and moderate categories), using FV offers superior supplementary details. Conversely, RGB+BEV proves superior for the challenging hard category comprising small and heavily occluded objects. This is attributed to its top-down perspective, eliminating occlusions inherent in other views. Incorporating BEV data significantly aids detection, especially in occlusion-heavy scenarios. As depicted in Table 1, when concatenating the BEV and FV modalities to utilize them produces optimal detection results across all categories using the benefits of both BEV and FV information.

Table 1. Average Precision (AP%) of TransfuseNet with respect to input data and late fusion operator.

Input Data	Late Fusion Operator	2D AP (%)		
		Easy	*Moderate*	*Hard*
RGB	No Fusion	87.08	83.13	77.07
RGB + BEV	Add	91.58	85.49	80.28
	Mul	92.12	86.21	82.19
	MFB	92.64	87.24	85.96
	MCF(ours)	92.81	88.51	86.15
RGB + FV	Add	92.19	86.93	79.94
	Mul	92.53	87.13	82.04
	MFB	92.77	88.13	85.57
	MCF(ours)	93.02	89.12	85.26
RGB + BEV + FV	Add	93.85	89.93	83.17
	Mul	94.39	90.78	83.96
	MFB	96.27	94.72	89.92
	MCF (ours)	**97.53**	**94.92**	**91.65**

Table 2. Performance evaluation of TransfuseNet using Average Precision (AP%). The table results are based on inputs of RGB images and concatenated BEV and FV representations.

Fusion operator	2D AP (%)			BEV AP (%)		
	Easy	*Moderate*	*Hard*	*Easy*	*Moderate*	*Hard*
Add	93.85	89.93	83.17	86.87	80.27	74.83
Mul	94.39	90.78	83.96	90.46	84.03	78.77
MFB	96.27	94.72	89.92	94.24	89.16	85.03
MCF (ours)	**97.53**	**94.92**	**91.65**	**94.93**	**91.02**	**87.12**

Different Fusion Operators. In this analysis, focusing on a singular input data type, we examine the impacts of varying late fusion operators. Special attention is paid to a direct quantitative comparison between our newly proposed MCF and the established MFB methods. As depicted in Table 1, our findings indicate that learnable fusion methods outclass their non-learnable counterparts. This performance gap is particularly noticeable when detecting smaller and more complex objects, designated as the 'hard' category in our case. Within this category, the shortcomings of non-learnable fusion methods become significantly more apparent, emphasizing the benefits of learnable approaches in challenging object detection scenarios. As shown in Table 2, considering the concatenation of RGB, BEV, and FV as input data, MCF achieves superior performance compared to MFB, registering gains of +1.26, +0.20 and +1.73% points for the

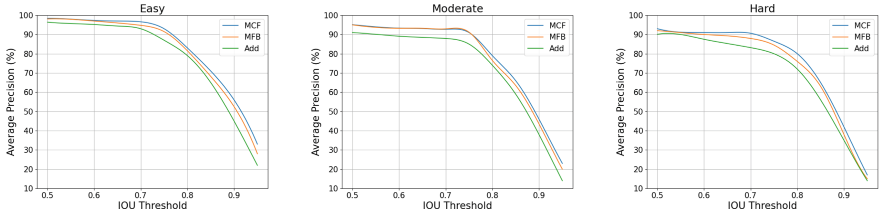


Fig. 4. TransfuseNet 2D Average Precision (AP%) vs. IoU evaluated using MCF, MFB, and element-wise addition fusion operators. Columns from left to right indicate results for easy, moderate, and hard categories.

easy, moderate hard categories, respectively, in 2D object detection. We further explored the impact of late fusion operators on BEV detection, as detailed in Table 2. In the BEV detection task, MCF surpasses MFB with significant margins of +0.69, +1.86, and +2.09% points, respectively. These improvements are primarily due to the MCF's employment of the priority gate which is sigmoid layer, which effectively identifies and amplifies the impact of more significant features within the fusion process.

Within the scope of non-learnable fusion methods, element-wise multiplication consistently outperforms addition across all evaluated scenarios. This superior efficacy of element-wise multiplication can be attributed to its inherent capability to amplify prominent features for transmission to subsequent layers, while simultaneously suppressing less significant features. These results collectively confirm the efficacy and superiority of our MCF fusion method over all studied fusion operators, specifically MFB, substantiating its applicability in both 2D and BEV object detection tasks.

In Table 4, we present the AP versus IoU performance of TransfuseNet, which employs various fusion operators, including MCF, MFB, and element-wise addition. These evaluations are conducted across the KITTI benchmark's three difficulty categories: easy, moderate, and hard. Our analysis reveals that learnable fusion operators consistently outperform the simpler element-wise addition approach, underscoring the importance of learnable fusion mechanisms in object detection. While MCF and MFB yield comparable performance at lower IoU thresholds, MCF exhibits superior performance at higher IoU values. This suggests that MCF generates predictions with greater confidence, rendering it a more reliable choice than MFB and element-wise addition.

State-of-the-Art Comparisons. As shown in Table 3, TransfuseNet, which utilizes our novel learnable fusion operator MCF, outperforms state-of-the-art networks in the BEV, easy, and moderate categories of 2D object detection. It achieves this with significantly less inference time compared to all other methods, demonstrating the network's efficiency—an essential factor for safe autonomous driving. Furthermore, as Table 4 reveals, TransfuseNet is three times faster than

Table 3. Evaluation results on KITTI 2D and BEV object detection benchmark (car). We evaluated TransfuseNet against the latest state-of-the-art results on the test set, using mean Average Precision measured at 40 recall positions for comparison. 'L' and 'I' represents LiDAR and Image, respectively. The best results appear in bold.

Method	Input		Time	2D AP (%)			BEV AP (%)		
	L	I	(ms)	*Easy*	*Moderate*	*Hard*	*Easy*	*Moderate*	*Hard*
OMNI3D[2]	-	✓	50	95.78	92.72	84.81	31.70	21.20	18.43
MonoNeRD[30]	-	✓	-	94.60	86.89	77.23	31.13	23.46	20.97
NeurOCS[19]	-	✓	100	96.39	91.08	81.20	32.27	24.49	20.89
IA-SSD[38]	✓	-	30	96.26	93.54	88.49	92.79	89.33	84.35
BtcDet[31]	✓	-	80	96.23	93.47	88.55	92.81	89.34	84.55
Pointpillars[15]	✓	-	30	94.00	91.19	88.17	90.07	86.56	82.81
PointRCNN[23]	✓	-	100	95.92	91.90	87.11	92.13	87.39	82.72
SVGA-Net[12]	✓	-	30	96.05	94.67	91.36	92.07	89.88	85.59
CAT-Det[37]	✓	✓	60	95.97	94.71	**92.07**	92.59	90.07	85.82
3D-CVF [34]	✓	✓	60	96.87	93.36	86.11	93.52	89.56	82.45
EPNet[13]	✓	✓	100	96.25	94.44	89.99	94.22	88.47	83.69
STD[33]	✓	✓	80	96.14	93.22	90.53	94.74	89.19	86.42
M3DETR[11]	✓	✓	180	97.39	94.83	91.10	94.41	90.37	85.98
TransfuseNet w/ MFB	✓	✓	-	96.27	94.72	89.92	94.24	89.16	85.03
TransfuseNet w/ MCF	✓	✓	**20**	**97.53**	**94.92**	91.65	**94.93**	**91.02**	**87.12**

Table 4. Comparative analysis of the number of parameters and inference time, evaluated on an NVIDIA GeForce RTX 3090 GPU with batch size 1.

Methods	CAT-Det [37]	M3DETR [11]	BtcDet [31]	TransfuseNet w/ MCF
# Param.	30M	76M	35M	**7M**
Inference (ms)	60	180	80	**20**

CAT-Det and nine times faster than M3DETR, both of which are transformer-based methods, while also having considerably fewer parameters.

4.5 Qualitative Results

Table 5 shows the advantages of LiDAR for 2D object detection. TransfuseNet effectively detects small and occluded objects, a task challenging for RGB-only models. This underscores the importance of incorporating LiDAR data for enhanced 2D object detection. Figure 6 further demonstrates the robustness of TransfuseNet; the network correctly identifies objects even when they are not labelled as ground truth.

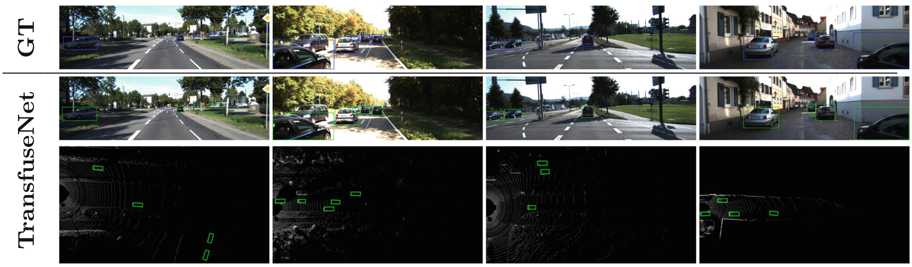


Fig. 5. Qualitative detection results of our TransfuseNet on KITTI validation samples. Green and blue bounding boxes are true positive detection and ground truth, respectively. (Color figure online)

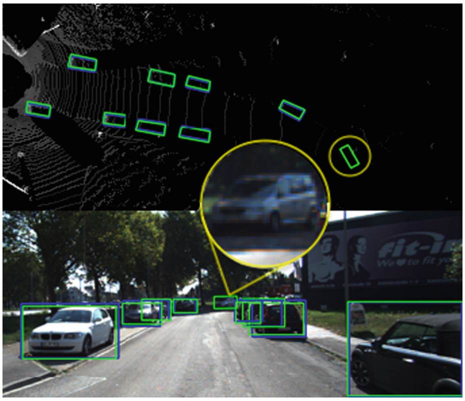


Fig. 6. Sample from the KITTI dataset illustrating the capability of our network to accurately detect an object despite incorrect annotation. The green bounding box indicates true positive detection, while the blue bounding box represents ground truth.

Table 5. Evaluating the impact of the number of transformer blocks employed in TransfuseNet.

# Transformer block	2D AP (%)		
	Easy	*Moderate*	*Hard*
0	87.97	84.92	81.67
1	95.75	90.93	85.03
2	**97.53**	**94.92**	**91.65**
3	96.88	94.53	89.94

Table 6. The effectiveness of different Mid-level fusion operator with two blocks utilized in TransfuseNet w/ MCF as late fusion operator.

Mid-level fusion operator	2D AP (%)		
	Easy	*Moderate*	*Hard*
No Mid-fusion	87.97	84.92	81.67
Add	88.71	86.52	81.79
Mul	89.93	87.67	83.41
Transformer	**97.53**	**94.92**	**91.65**

4.6 Ablation Study

In this section, we conducted ablation studies on various facets, including input data types, backbone models, fusion techniques, and Transformer structure.

Initially, we evaluated the efficacy of multi-view features across various input modalities and late fusion operators. These results are encapsulated in Table 1. Significantly, our results demonstrate that the integration of RGB, BEV, and FV inputs consistently outperforms other combinations, irrespective of the late fusion operator used. Moreover, our proposed MCF method exhibited superior performance compared to the alternative fusion techniques in all scenarios.

We assessed the impact of varying the number of Transformer blocks within TransfuseNet. Table 5 highlights that optimal performance is achieved with two Transformer blocks. The absence of Transformer blocks, and hence mid-level fusion, significantly degrades results, underscoring the efficacy of Transformer blocks in enhancing mid-level fusion. The diminishing returns observed with three blocks suggest that an increased parameter count may impede performance. Furthermore, Table 6 shows the superior effectiveness of employing Transformers as mid-level fusion operators, compared to addition or multiplication methods.

Finally, as indicated in Table 7, we compared our model against various parameters such as the number of attention heads and layers, and different backbone models. In our default configuration, we use two Transformer layers, eight attention layers, four attention heads and ConvMixer for BEV and RGB feature extraction. In selecting a feature extractor, we chose ConMixer as our primary backbone due to its superior accuracy over alternatives like VGG-16 and ResNet-34. Notably, ConMixer offers a simpler structure and requires significantly fewer parameters, approximately only one-tenth those of ResNet-34.

Table 7. Ablation study of 2D object detection. Comparison of different model structures' results on the KITTI validation set.

Network Parameter	Value	2D AP (%)		
		Easy	*Moderate*	*Hard*
Attention layer	2	95.76	91.25	86.61
	4	95.62	93.93	86.92
	6	96.49	93.01	88.88
Attention head	2	96.12	93.48	87.75
Backbone	VGG-16	96.24	91.65	90.43
	ResNet-34	96.97	93.26	88.54
Default Config	-	**97.53**	**94.92**	**91.65**

5 Conclusion

In this work, we introduced TransfuseNet, an innovative end-to-end, two-stage object detection framework. The primary aim of TransfuseNet is to provide a conceptually simple, fast, yet accurate fusion method that is useful for making quick decisions in critical autonomous driving scenarios, such as collision avoidance. TransfuseNet effectively leverages Transformer-based fusion, capitalizing on the synergies between LiDAR and camera modalities for 2D object detection. While Transformers facilitate mid-level fusion, an extensive examination of fusion operators has offered valuable insights for the late fusion stage. Notably, our proposed learnable fusion technique resulted in a significant improvement in both 2D and BEV detection performance. This progress contributes to the development of fusion methods and highlights potential areas for future research in multi-modal object detection.

References

1. Bai, X., et al.: Transfusion: robust lidar-camera fusion for 3d object detection with transformers. In: Proceedings of the IEEE/CVF Conference on Computer Vision And Pattern Recognition, pp. 1090–1099 (2022)
2. Brazil, G., Kumar, A., Straub, J., Ravi, N., Johnson, J., Gkioxari, G.: Omni3d: a large benchmark and model for 3d object detection in the wild. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 13154–13164 (2023)
3. Chai, Y., et al.: To the point: efficient 3d object detection in the range image with graph convolution kernels. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 16000–16009 (2021)
4. Chen, Y., Li, Y., Zhang, X., Sun, J., Jia, J.: Focal sparse convolutional networks for 3d object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 5428–5437 (2022)
5. Dosovitskiy, A., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
6. El Ahmar, W., et al.: Enhanced thermal-rgb fusion for robust object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 365–374 (2023)
7. Fan, L., Xiong, X., Wang, F., Wang, N., Zhang, Z.: Rangedet: in defense of range view for lidar-based 3d object detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 2918–2927 (2021)
8. Fu, C.Y., Liu, W., Ranga, A., Tyagi, A., Berg, A.C.: Dssd: Deconvolutional single shot detector. arXiv preprint arXiv:1701.06659 (2017)
9. Geiger, A., Lenz, P., Urtasun, R.: Are we ready for autonomous driving? the kitti vision benchmark suite. In: 2012 IEEE Conference on Computer Vision and Pattern Recognition, pp. 3354–3361. IEEE (2012)
10. Girshick, R.: Fast r-cnn. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 1440–1448 (2015)
11. Guan, T., et al.: M3detr: multi-representation, multi-scale, mutual-relation 3d object detection with transformers. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pp. 772–782 (2022)
12. He, Q., Wang, Z., Zeng, H., Zeng, Y., Liu, Y.: Svga-net: sparse voxel-graph attention network for 3d object detection from point clouds. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 36, pp. 870–878 (2022)
13. Huang, T., Liu, Z., Chen, X., Bai, X.: EPNet: Enhancing Point Features with Image Semantics for 3D Object Detection. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J.-M. (eds.) ECCV 2020. LNCS, vol. 12360, pp. 35–52. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-58555-6_3
14. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
15. Lang, A.H., Vora, S., Caesar, H., Zhou, L., Yang, J., Beijbom, O.: Pointpillars: fast encoders for object detection from point clouds. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 12697–12705 (2019)
16. Li, C., et al.: Yolov6: A single-stage object detection framework for industrial applications. arXiv preprint arXiv:2209.02976 (2022)
17. , Li, Y., et al.: Deepfusion: Lidar-camera deep fusion for multi-modal 3d object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 17182–17191 (2022)

18. Liu, Z., et al.: Bevfusion: multi-task multi-sensor fusion with unified bird's-eye view representation. In: 2023 IEEE International Conference on Robotics and Automation (ICRA), pp. 2774–2781. IEEE (2023)
19. Min, Z., Zhuang, B., Schulter, S., Liu, B., Dunn, E., Chandraker, M.: Neurocs: neural nocs supervision for monocular 3d object localization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 21404–21414 (2023)
20. Pasandi, M.M., Liu, T., Massoud, Y., Laganière, R.: Sensor fusion operators for multimodal 2d object detection. In: International Symposium on Visual Computing, pp. 184–195. Springer (2022)
21. Qi, D., Su, L., Song, J., Cui, E., Bharti, T., Sacheti, A.: Imagebert: Cross-modal pre-training with large-scale weak-supervised image-text data. arXiv preprint arXiv:2001.07966 (2020)
22. Ren, S., He, K., Girshick, R., Sun, J.: Faster r-cnn: tDowards real-time object detection with region proposal networks. Advances in Neural Information Processing Systems **28** (2015)
23. Shi, S., Wang, X., Li, H.: Pointrcnn: 3d object proposal generation and detection from point cloud. In: Proceedings of the IEEE/CVF Conference on Computer Vision And Pattern Recognition, pp. 770–779 (2019)
24. Sun, P., et al.: Scalability in perception for autonomous driving: Waymo open dataset. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 2446–2454 (2020)
25. Sun, P., et al.: Rsn: range sparse net for efficient, accurate lidar 3d object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 5725–5734 (2021)
26. Trockman, A., Kolter, J.Z.: Patches are all you need? arXiv preprint arXiv:2201.09792 (2022)
27. Vaswani, A., et al.: Attention is all you need. In: Advances in Neural Information Processing Systems, vol. 30 (2017)
28. Wang, C.Y., Bochkovskiy, A., Liao, H.Y.M.: Yolov7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 7464–7475 (2023)
29. Wang, C., Ma, C., Zhu, M., Yang, X.: Pointaugmenting: Cross-modal augmentation for 3d object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 11794–11803 (2021)
30. Xu, J., et al.: Mononerd: Nerf-like representations for monocular 3d object detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 6814–6824 (2023)
31. Xu, Q., Zhong, Y., Neumann, U.: Behind the curtain: Learning occluded shapes for 3d object detection. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 36, pp. 2893–2901 (2022)
32. Yang, H., et al.: Pvt-ssd: Single-stage 3d object detector with point-voxel transformer. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 13476–13487 (2023)
33. Yang, Z., Sun, Y., Liu, S., Shen, X., Jia, J.: Std: Sparse-to-dense 3d object detector for point cloud. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 1951–1960 (2019)
34. Yoo, J.H., Kim, Y., Kim, J., Choi, J.W.: 3D-CVF: Generating Joint Camera and LiDAR Features Using Cross-view Spatial Feature Fusion for 3D Object Detection.

In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J.-M. (eds.) ECCV 2020. LNCS, vol. 12372, pp. 720–736. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-58583-9_43

35. Yu, C., Gao, C., Wang, J., Yu, G., Shen, C., Sang, N.: Bisenet v2: bilateral network with guided aggregation for real-time semantic segmentation. Int. J. Comput. Vision **129**, 3051–3068 (2021)
36. Yu, Z., Yu, J., Fan, J., Tao, D.: Multi-modal factorized bilinear pooling with co-attention learning for visual question answering. In: Proceedings of the IEEE International Conference On Computer Vision, pp. 1821–1830 (2017)
37. Zhang, Y., Chen, J., Huang, D.: Cat-det: contrastively augmented transformer for multi-modal 3d object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 908–917 (2022)
38. Zhang, Y., Hu, Q., Xu, G., Ma, Y., Wan, J., Guo, Y.: Not all points are equal: learning highly efficient point-based detectors for 3d lidar point clouds. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 18953–18962 (2022)
39. Zhou, C., Zhang, Y., Chen, J., Huang, D.: Octr: Octree-based transformer for 3d object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 5166–5175 (2023)

Author Index

A
Adhikarla, Eashan 260
Aithal, Manjushree 260

B
Babazaki, Yasunori 229
Bagci, Ulas 354
Banerjee, Ayan 17
Bansal, Ankur 398
Basak, Suvramalya 179
Bhattacharya, Rohit 65
Bravo-Rocca, Gusseppe 337
Burghouts, Gertjan J. 291

C
Chen, Chi 98
Cheng, Yu 307
Choi, Jinwoo 116

D
Davison, Brian D. 260
Dey, Avijit 164
Dholakia, Ajay 337
Doan, Khoa D. 276

E
Ellison, David 337

F
Fang, Yi 1
Feng, Miao 212
Fu, Hao 413

G
Gautam, Anjali 179
Geißler, Daniel 133
Ghosh, Banibrata 276
Guitart, Jordi 337
Gupta, Sandeep K. S. 17

H
Hao, Jie 366
Hao, Yu 1
Harikumar, Haripriya 276
He, Jinhong 307
Hendrix, Rutger 354
Hooshyaripour, Niloofar 445
Hu, Jiannan 148
Huang, Hao 1
Huang, Sirong 51

J
Jakhetiya, Vinit 164, 398
Jia, Penghao 307

K
Kapadia, Mubbasir 382
Kapoor, Meghna 398
Kirtas, Manos 322
Kittler, Josef 36, 148

L
Lamrani, Imane 17
Lee, WonSook 445
Li, Bo 307
Li, Danrui 382
Liao, Tongkui 413
Liu, Guozhang 413
Liu, Mengxi 133
Liu, Peini 337
Liu, Qingze Tony 382
Liu, Ting 413
Liu, Xuexin 366
Liu, Yiyun 1
Lukowicz, Paul 133
Lv, Kewei 98

M
Madhusudhana, Nikhil Ambha 260
Maiti, Abhishek 116

A. Antonacopoulos et al. (Eds.): ICPR 2024, LNCS 15329, pp. 461–462, 2025.
https://doi.org/10.1007/978-3-031-78110-0

Majumdar, Puspita 65
Manzoor, Ayesha 81
Meunier, Jean 212

N
Nicholson, John 260
Ning, Aoxiang 307

P
Palazzo, Simone 354
Pang, Tao 413
Passalis, Nikolaos 322
Pavlovic, Vladimir 382
Peng, Silong 245
Prajesh, Balraj 65
Proietto Salanitri, Federica 354

R
Rajan, Deepu 196
Rana, Santu 276
Rattani, Ajita 81
Ray, Lala Shakti Swarup 133

S
Sadeghian, Reza 445
Sharma, Raghav 65
Shen, Peisong 98
Shibata, Takashi 229
Shin, Jongmin 116
Singh, Himanshu 164
Sohn, Samuel S. 382
Spampinato, Concetto 354
Subudhi, Badri Narayan 164, 398
Suh, Sungho 133
Sun, Lichao 260

T
Takahashi, Toru 229
Tefas, Anastasios 322

V
Vaishy, Aditya 179
van Rooij, Sabina B. 291
Venkatesh, Svetha 276
VidalMata, Rosaura G. 260

W
Wang, Tao 413
Wen, Congcong 1
Wu, Cong 148
Wu, Si 429
Wu, Xiao-Jun 36, 148
Wu, Ziyu 245

X
Xu, Tianyang 36, 148

Y
Yang, Guangxing 413
Yihui, Qiu 196
Yoon, Sejong 382
Yuan, Mengke 413
Yuan, Shuaihang 1

Z
Zhang, Kai 260
Zhang, Ruoqi 98
Zhang, Yinhao 36
Zhang, Yunfei 429
Zhao, Shaochuan 36
Zheng, Huicheng 51
Zhong, Wenlong 429
Zhou, Bo 133
Zou, Yuliang 116
Zou, Zhuojun 366