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

LNCS 15313

Pattern Recognition

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

Lecture Notes in Computer Science **15313**

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 XIII

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 978-3-031-78200-8 ISBN 978-3-031-78201-5 (eBook)
https://doi.org/10.1007/978-3-031-78201-5

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

President's Address

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

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

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

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

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

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

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

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

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

September 2024

Arjan Kuijper
President of the IAPR

Preface

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

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

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

In total, 945 papers (669 from round-1 and 276 from round-2) were accepted for presentation, resulting in an acceptance rate of 44.26%, which is consistent with previous ICPR events. 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	Da-Han Wang
Mingxin Huang	Yang Xue
Lei Kang	Wentao Yang
Wenhui Liao	Jiaxin Zhang
Yuliang Liu	Yiwu Zhong
Yongxin Shi	

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 XIII

A Single Source Generalization Model via Spatial Amplitude Perturbation and Sensitivity Guidance for Colored Medical Image Segmentation

Zeyuan Yang and Chunyan Yu(✉)

College of Computer and Data Science/College of Software, Fuzhou University, Fuzhou 350108, China
therica@fzu.edu.cn

Abstract. For medical images, domain shift is a very common phenomenon. To address this issue, researchers have proposed unsupervised domain adaptation and multi-source domain generalization. However, these methods are sometimes impractical for clinical applications since they need multi-domain data. To this end, single-source domain generalization has been further proposed. However, most single-source domain generalization methods are designed for grayscale medical images, making them unsuitable for color images such as fundus images. In this paper, we first propose a novel and effective Fourier transform-based data augmentation method for single-source domain color medical images, named spatial amplitude perturbation module (SAPM). The SAPM uses different Gaussian distributions to perturb different regions of the amplitude map obtained by FFT decomposition, thereby avoiding the need for information from other domains and ensuring the diversity of the augmented images. Then, we use feature sensitivity to guide the network to learn domain-invariant features, which can suppress feature channels sensitive to domain shift and emphasize feature channels insensitive to domain shift. We evaluate our method on a multi-domain fundus segmentation benchmark, and the results demonstrate the effectiveness of our proposed method.

Keywords: Domain generalization · Data augmentation · Medical image segmentation

1 Introduction

In recent years, deep learning has made significant achievements in medical image segmentation [10,17,19]. However, such achievements are based on the assumption that the training data and testing data come from same domain. Unfortunately, such an assumption often does not hold true in clinical applications due to the variations in environmental factors, patient or disease severity during

A. Antonacopoulos et al. (Eds.): ICPR 2024, LNCS 15313, pp. 1–16, 2025.
https://doi.org/10.1007/978-3-031-78201-5_1

data acquisition, which often exhibit variations in the field of view, appearance, and image quality. As a result, the performance of a model that is trained on a source domain will drop sharply while the model processes data from unseen domains, which severely impedes the application of deep learning models in clinical settings [13].

To address this issue, researchers first proposed unsupervised domain adaption (UDA) [25]. UDA methods learn the network on annotated source domain images and unlabeled target domain images in order to make the network have good generalization performance on target domains. Therefore, the goal of the UDA is to narrow the domain gap between the source domain and the target domain, which makes it necessary for the UDA to acquire the target domain data in advance. However, such a requirement is often infeasible in clinical applications for medical images segmentation.

To solve the limitation in UDA, researchers further propose domain generalization (DG) [22]. In comparison to UDA, DG does not need to access target domains, which means that the trained model can directly apply to unseen target domains. Classic DG setting assumes that access to multiple source domains is feasible during training, which namely multi-source domain generalization (multi-DG). Multi-DG methods can be mainly divided into data augmentation, domain alignment, and meta-learning [31]. However, collecting and labelling data from multiple domains is a time-consuming and labour-intensive process, especially for medical image data, which seriously hinders the clinical application of domain generalization in medical image segmentation.

A more challenging yet realistic DG scenario is the single-source domain generalization (single-DG), wherein the model is trained on labelled data from one single source domain and subsequently applied to unseen domains [16]. The primary challenge of single-DG is the constrained diversity of samples, which makes the trained model susceptible to overfitting on the single source domain. Therefore, one of the most straightforward approaches in single-DG is to expand the diversity of single-source domain data through data augmentation. There has been much research on single-source medical image data augmentation [14,20, 28,34], but these works are predominantly tailored for grayscale medical images like CT or MRI scans, making them unsuitable for color medical images.

For color medical images like fundus images [23], most domain generalization studies resort to data augmentation techniques based on Fourier transformation [27,33]. These methodologies typically first decompose images via fast Fourier transformation (FFT) to acquire amplitude maps and subsequently exchange the low-frequency parts of the amplitude maps from different domains. Finally, the augmented images are obtained through inverse fast Fourier transformation (IFFT). Regrettably, these Fourier-based methods are not feasible for single-DG settings due to their reliance on information from other domains. To make the Fourier transformation applicable in the single-DG setting, it is imperative to execute appropriate operations on the decomposed amplitude map. The most intuitive approach is to apply Gaussian perturbation on the amplitude map. However, owing to significant differences in the distribution of values in differ-

ent parts of the amplitude map, simply perturbing the entire amplitude map using the same Gaussian distribution is not a reasonable approach. To this end, we propose a simple and effective Fourier transformation-based augmentation method, named spatial amplitude perturbation module (SAPM), which adopts different Gaussian perturbations for different locations on the amplitude map to simulate domain shift better.

Besides, since the domain shift caused by Gaussian perturbation is limited, if origin images and augmented images are only fed into the network for training without explicitly guiding the network to learn domain-invariant features, the network may learn domain-shared information rather than the domain-invariant information [12]. The domain-shared information is usually shared information between several domains, rather than domain-invariant information common to all domains, which means that domain-shared information may sometimes not work well for other unseen domains. Thus, we will also apply a method called feature sensitivity guidance to explicitly guide the network to learn domain-invariant features. This method can suppress domain-sensitive features and emphasize domain-insensitive features, which have been proven effective in [26].

In general, our contributions are summarized as follows:

- We propose a novel Fourier-based data augmentation method for color medical images that can significantly expand data diversity without using other domain information.
- We employ feature sensitivity to explicitly guide the network to learn domain-invariant features to promote the generalization ability of our color medical images segmentation network.
- We compare our method with six recent single-DG methods on the single-domain generalization tasks tailored to fundus OD/OC segmentation. Extensive experimental results show that our proposed method is superior to the compared methods.

2 Related Works

2.1 Single Source Domain Generalization

Single-source domain generalization is a more challenging task than multiple-source domain generalization due to the lack of diversity of the training set, which also makes it more realistic than multi-DG. The most popular method to resolve single-DG is data augmentation, which imitates domain shift by generating pseudo domains different from the source domain. Several methods have been designed for augmentation in single source domain generalization tasks through different strategies [9,12,32]. As for the medical image, inspired by [35], many methods change the pixel intensity of the source image by using monotonic nonlinear functions, such as Bezier curves, to expand the diversity of data. [14] first, transformed the origin image twice to obtain two completely different augmented images, and then blended the two images in a spatial-variables manner

to remove the spurious correlations and constrain the model to make consistent predictions of these two blended images, which finally makes the model not affected by images' appearance and spurious correlations. [20] observed that the distribution of pixel intensity in different mask regions is various. Thus, they augmented the source images by using their masks as guides instead of simply performing the full image-level transformation. However, the above methods are usually applied to grayscale images and not to RGB images, which means that these methods are not suitable for color medical images such as fundus images. [11] obtained different frequency views through different Gaussian filters, then exchanged random portions of these views to obtain augmented images and learned generalizable context-aware representations through a self-supervised task, but this approach is only applicable to tasks like vessel segmentation.

Also, some single-DG methods focus on learning domain-invariant features by introducing approaches such as the instance normalization layers [21] or feature whitening transformation to remove domain-specific information [3,18].

2.2 Data Augmentation for Domain Generalization

Data augmentation is one of the most prevailing methods in domain generalization, and it is common practice for such methods to perform a stylistic transformation on the source domain image. [24] generated diverse images through adversarial training, and [29] proposed random convolution to perturb original images. Recently, numerous studies have used the Fourier transformation to generate new images [15,27,30], which are based on a widely accepted notion that the phase of an image contains semantic information while the amplitude of an image contains style information. Specifically, these methods first decompose the images into phases and amplitudes through fast Fourier transformation (FFT), and then they exchange the amplitudes of the images from different domains. Finally, the original phase and the exchanged amplitudes are subjected to inverse fast Fourier transformation (IFFT) to acquire augmented images. The method is also widely applied in medical images. For example, [33] obtained new images by exchanging low-frequency components of fundus images from different domains. However, existing Fourier-based approaches are only applicable in the case of multiple source domains.

3 Method

3.1 Overview

Let the source domain be denoted by $D^S = \{x_i^S, y_i^S\}_{i=1}^{N_S}$, where x_i^S is the i-th source domain image, and y_i^S is the corresponding mask. Our goal is to train a robust segmentation model $f_\theta : x \rightarrow y$ on D^S that can generalize well to the unseen domain images.

The method we proposed mainly consists of a spatial amplitude perturbation module (SAPM), a sensitivity guidance module and a segmentation backbone.

The overall structure of our model is shown in Fig. 1. For image x_j^S, our SAPM generates a augmented images x_j^A, which have the same semantic information with some domain shift, the subscript j presents the j-th augmented image. Then, we feed the origin image and augmented image into the segmentation network to get the prediction $\widetilde{y}$. To learn more robust feature representation, feature sensitivity will be used to guide the network to suppress shallow feature channels that are sensitive to domain shift.

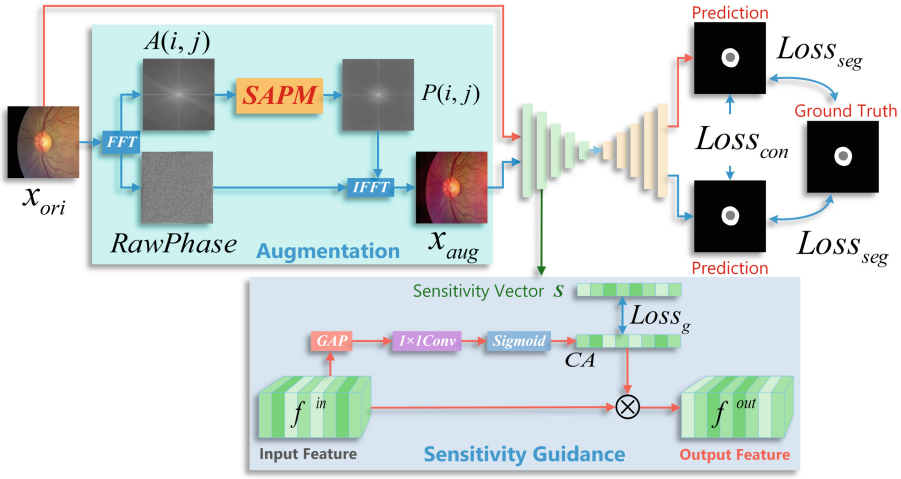


Fig. 1. Our model's overview. The origin image x_{ori} first passes through the SAPM to obtain augmented image x_{aug}, where $A(i,j)$ and $P(i,j)$ are the original and augmented amplitude maps, respectively. Then, the two images are input into network to extract features and calculate the sensitivity vector S, and domain-sensitive features will be suppressed by the guidance of S. Finally, the features f_{out} will be fed into subsequent layers to obtain predictions. Segmentation loss $Loss_{seg}$, consistency loss $Loss_{con}$ and guidance loss $Loss_g$ are employed to optimize our network.

3.2 Spatial Amplitude Perturbation Module (SAPM)

Prior works about Fourier augmentation methods are based on a consensus that the phase component of the Fourier spectrum mainly retains semantic information, and the amplitude contains style features. Hence, these methods transform the image by exchanging the amplitude signals of the images in different domains. However, in the single-DG setting, we are unable to obtain information from other domains, so we choose to directly perturb the amplitude signal reasonably.

An observation inspired the module. We decompose the source image through FFT to get its amplitude, then shift the low-frequency components to the center

and visualize it, as shown in Fig. 2. In the image, the center is brighter than the surrounding areas, and the brightness fades outward in a circular pattern. It should be noted that the relative pixel values in the image are calculated by applying a logarithmic function to the original amplitude value. Next, we view the raw values of the amplitude values and find that the values of the low frequency in the center are much higher than in the other regions. Typically, values in the centermost area are in the thousands to hundreds of thousands, while values in the surrounding region decay rapidly to tens to hundreds, and values in the outermost region are usually in the single digits.

In general, the perturbation for an image signal is to multiply the original value by a scaled value obtained by sampling in a Gaussian distribution, and we follow this practice for the perturbation of the amplitude map. However, due to the significant differences in the amplitude values of the different frequencies, simply scaling all the values with a same distribution would result in a limited augmentation. Combined with the above observations, we believe that the degree of perturbation should be small for the central low-frequency values, while the degree of perturbation can be gradually increased for the surrounding values, i.e., the larger the value, the smaller the degree of perturbation, and vice versa.

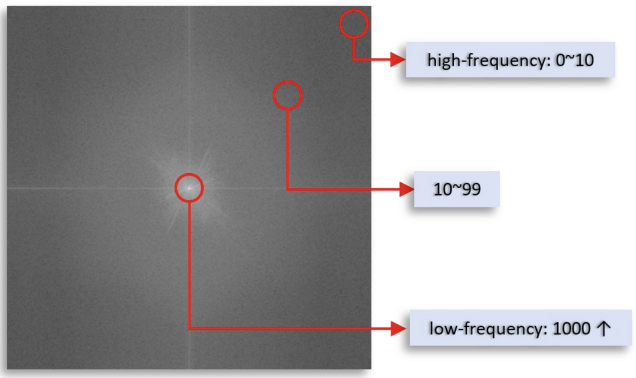


Fig. 2. Our observation of amplitude map (values processed by the log function). It can be seen that values in the low-frequency region are much higher than those in the high-frequency region.

Based on the above views, we propose the Spatial Amplitude Perturbation Module (SAPM). We first decompose the original image into amplitude and phase maps through FFT. Denoted the amplitude map as $A(i, j)$, the perturbed amplitude map as $P(i, j)$, where (i, j) is the coordinates of the points in amplitude map and $(0, 0)$ is the center point coordinate, $i \in \{-\frac{H}{2}, \frac{H}{2}\}, j \in \{-\frac{W}{2}, \frac{W}{2}\}$, where H and W represent the height and width of the image, respectively.

For the center point $(0, 0)$, its perturbation scaling values $\mu(0, 0)$ can be obtained by sampling from a Gaussian distribution with a mean of 1 and a

variance offset of β, β usually is set as 0.5, and then the original value $A(0,0)$ is multiplied by the perturbation scaling value to get the perturbed value $P(0,0)$, which can be expressed as:

$$P(0,0) = A(0,0) * \mu(0,0), \quad \mu(0,0) \sim N(1,\beta) \tag{1}$$

For other points, we use the spatial distance between the point and the center point to determine the variance value of the Gaussian distribution that is used for sampling perturbation scaling value. The spatial distance $D(i,j)$ can be acquired by following formula:

$$D(i,j) = \frac{i^2 + j^2}{max(i^2 + j^2)} \tag{2}$$

where $max(i^2 + j^2)$ is the maximum distance from the center point, which is divided to obtain the normalized distance. Then the variance $\sigma^2_{i,j}$ is formulated as:

$$\sigma^2_{i,j} = \alpha * D(i,j) + \beta \tag{3}$$

where α is the distance coefficient and is set as 6, β is the variance offset in Eq. 1. Finally, the perturbed amplitude map can be represented as:

$$P(i,j) = A(i,j) * \mu(i,j), \quad \mu(i,j) \sim N(1,\sigma^2_{i,j}) \tag{4}$$

After obtaining the perturbed amplitude map $P(i,j)$, combine it with raw phase map through IFFT to get the augmented image:

$$x_{aug} = IFFT(P(i,j), RawPhase) \tag{5}$$

3.3 Features Sensitivity Guidance

After obtaining the augmented image, we feed it along with the raw image to the subsequent segmentation network. Although SAPM can simulate domain shift and greatly expand data diversity, the ability of Gaussian perturbations is limited, which may result in the network learning more about domain-shared features. For this purpose, we apply a simple and effective method to guide the network to learn domain-invariant features.

Previous works on single-DG [8,26] have shown that shallow features in different channels are differently sensitive to domain shift. Hence, we expect the segmentation network to have the ability to learn channels that are greatly affected by domain shift and suppress them.

The method we performed is called feature sensitivity guidance, which learns domain-invariant features from the perspective of feature channels. The sensitivity guidance guides the network to focus on domain-invariant feature channels via SENet [6].

Specifically, we first extract the low-level features f_{ori} and f_{aug} from the raw image x_{ori} and augmented image x_{aug} and then subtract them to obtain the feature differences vector FD, which can be described as:

$$FD \in R^{B\times C\times 1\times 1} = GAP(\|f_{ori} - f_{aug}\|) \tag{6}$$

where $f_{ori}, f_{aug} \in R^{B\times C\times H\times W}$, and GAP is global average pooling. Then the sensitivity vector $S \in R^{B\times C\times H\times W}$ can be obtained by normalizing FD:

$$S = \frac{FD - FD_{min}}{FD_{max} - FD_{min}} \tag{7}$$

where FD_{min} is the minimum of FD and FD_{max} is the maximum of FD. The feature sensitivity vector S reflects how sensitive the features of different channels are to domain shift. A channel with a larger value means that it is more sensitive to domain shift and vice versa, those feature channels that are not sensitive to domain shift are considered as domain-invariant features.

Then f_{ori} and f_{aug} as input features f^{in} are send to SE module to get channel attentions CA and output features f^{out} respectively, which can be formulated as:

$$CA \in R^{B\times C\times 1\times 1} = \sigma(Conv(GAP(f^{in}))) \tag{8}$$

$$f^{out} = CA \oplus f^{in} \tag{9}$$

where GAP is the global average pooling, $Conv$ is a 1×1 convolution layer and σ refer to sigmoid function. We hope that the channel attention CA and the sensitivity vector S are negatively correlated because we want the network to pay more attention to channels that are not sensitive to domain shift. The loss between them is defined as guidance loss $Loss_g$:

$$Loss_g = \|log(CA)log(S) - 1\|_2 \tag{10}$$

The $Loss_g$ can constraint the attention CA close to 0 when sensitivity vector S close to 1 for each channel, which will guide the network to suppress channels sensitive to domain shift and emphasize channels insensitive to domain shift.

For the loss of segmentation task $Loss_{seg}$, we employ binary cross-entropy loss and dice loss. Meanwhile, we constrain the consistency between the output of the origin image and the output of the augmented image through consistency loss $Loss_{con}$, which can be represented as:

$$Loss_{con} = MSE(y_{ori} - y_{aug}) \tag{11}$$

where y_{ori}, y_{aug} are the network output of x_{ori}, x_{aug} that processed by sigmoid function. Thus, the total loss can be written as:

$$Loss_{total} = \lambda_1 * Loss_{seg} + \lambda_2 * Loss_{con} + \lambda_3 * Loss_g \tag{12}$$

where $\lambda_1, \lambda_2, \lambda_3$ are hyperparameters to balance each loss, set to 1, 0.5, and 0.5 respectively in our experiment.

4 Experiments and Results

4.1 Datasets and Evaluation Metrics

The dataset used in this paper is the RIGA+ dataset [7], which is first utilized to evaluate unsupervised domain adaption method in fundus image segmentation.

Specifically, the RIGA+ dataset contains annotated fundus images from five domains across two datasets, among which BinRushed and Magrabia are from the RIGA dataset [1], and Base1, Base2, and Base3 are from the MESSIDOR dataset [4]; every image is annotated by six ophthalmologists, we choose the first ophthalmologist's annotations as the masks. Data from different domains are captured by different devices or different medical institutes.

For the evaluation, we use BinRushed and Magrabia as source domains for training separately, then evaluate performance of models on images from the Base1, Base2, and Base3 domains. The dice similarity coefficient is employed to evaluate models' segmentation performance, and a higher Dice coefficient represents better performance.

4.2 Implementation Details

The origin images are center-cropped and resized to 512×512 and normalized to [0, 1], the batch size is set to 8. We employ a U-net based structure as the segmentation network for our method as well as for all the competing methods, and we employ a modified ResNet-34 [5] as the encoder. We implement our experiment with the PyTorch framework on one Nvidia RTX 3080 GPU with 10 GB memory and train the model for 200 epochs on each source domain. We also employ the SGD optimizer with an initial learning rate of 0.01 and a momentum of 0.99 to optimize our model. We utilize the average of the test results of the last 10 epochs as our method's final result because the network has converged by this time. We also report the standard deviation of the final result, which can illustrate the stability of the network.

4.3 Comparative Experiments

Table 1. Comparison results of using BinRushed as source domain

Methods	Base1		Base2		Base3		Average	
	OD_{std}	OC_{std}	OD_{std}	OC_{std}	OD_{std}	OC_{std}	OD_{std}	OC_{std}
w/o SDG	$89.77_{1.14}$	$75.71_{0.68}$	$79.32_{1.22}$	$68.39_{2.17}$	$87.03_{1.59}$	$76.27_{2.31}$	85.37	73.46
MaxStyle [2]	$93.18_{0.34}$	$82.33_{0.76}$	$87.97_{0.96}$	$75.89_{0.88}$	$93.07_{0.51}$	$83.21_{0.74}$	91.41	80.48
GIN-IPA [14]	$93.07_{0.77}$	$81.92_{1.34}$	$92.65_{0.92}$	$81.01_{0.98}$	$92.87_{0.71}$	$82.20_{1.01}$	92.86	81.71
ADS [28]	$93.51_{0.64}$	$80.25_{1.71}$	$93.18_{0.66}$	$82.03_{0.49}$	$92.79_{0.19}$	$81.93_{1.60}$	93.15	81.40
Dual-Norm [34]	$92.99_{1.89}$	$82.38_{0.92}$	$91.72_{0.52}$	$80.29_{1.22}$	$92.31_{0.44}$	$82.93_{0.87}$	92.34	81.87
SLAug [20]	$94.11_{0.39}$	$84.07_{0.95}$	$93.96_{0.36}$	$82.15_{1.21}$	$94.78_{0.43}$	$83.00_{0.27}$	94.28	83.07
CCSDG [8]	$\mathbf{96.07_{0.27}}$	$85.69_{0.31}$	$95.09_{0.24}$	$84.88_{0.48}$	$95.89_{0.09}$	$85.10_{0.19}$	95.68	85.22
Ours	$95.93_{0.09}$	$\mathbf{86.79_{0.16}}$	$\mathbf{96.01_{0.24}}$	$\mathbf{85.01_{0.12}}$	$\mathbf{95.96_{0.17}}$	$\mathbf{86.87_{0.21}}$	**95.96**	**86.22**

We first compare our method with one baseline, named 'w/o SDG' (*i.e.* training on source domains then testing on target domains without using single-DG

Table 2. Comparison results of using Magrabia as source domain

Methods	Base1		Base2		Base3		Average	
	OD_{std}	OC_{std}	OD_{std}	OC_{std}	OD_{std}	OC_{std}	OD_{std}	OC_{std}
w/o SDG	$87.20_{0.44}$	$75.81_{0.79}$	$83.63_{1.11}$	$74.98_{0.42}$	$88.93_{0.39}$	$79.86_{0.57}$	86.59	76.88
MaxStyle [2]	$90.33_{0.32}$	$79.63_{1.27}$	$89.88_{0.33}$	$79.82_{0.70}$	$91.39_{0.67}$	$81.97_{1.02}$	90.53	80.47
GIN-IPA [14]	$90.07_{0.74}$	$78.00_{1.07}$	$87.34_{0.92}$	$76.77_{1.83}$	90.350.27	81.211.34	89.25	78.66
ADS [28]	$90.92_{1.03}$	$78.78_{0.56}$	$90.11_{1.97}$	$79.19_{1.07}$	$91.26_{1.99}$	$80.55_{1.28}$	90.76	79.51
Dual-Norm [34]	$93.01_{0.30}$	$81.38_{0.56}$	$92.13_{0.39}$	$81.19_{0.42}$	$92.33_{0.36}$	$81.25_{0.64}$	92.49	81.27
SLAug [20]	$92.78_{0.87}$	$82.71_{1.18}$	$93.01_{0.36}$	$81.55_{0.31}$	$93.43_{0.26}$	$81.78_{0.54}$	93.07	82.01
CCSDG [8]	$95.22_{0.13}$	$85.69_{0.33}$	94.500.17	$85.03_{0.29}$	$94.62_{0.07}$	$86.40_{0.37}$	94.78	85.71
Ours	$\mathbf{95.37_{0.06}}$	$\mathbf{85.91_{0.19}}$	$\mathbf{95.93_{0.11}}$	$\mathbf{85.88_{0.13}}$	$\mathbf{95.64_{0.03}}$	$\mathbf{86.78_{0.17}}$	**95.65**	**86.19**

methods). Then, we choose six recent medical SDG methods to compare with our method, namely MaxStyle [2], GIN-IPA [14], ADS [28], Dual-Norm [34], SLAug [20], CCSDG [8]. As we mentioned in the introduction, most single-DG methods are designed for grayscale images, so most of these methods are also designed for grayscale images. MaxStyle [2] extended the style space by introducing noise and adversarial training. GIN-IPA [14] augmented images by global intensity non-linear augmentation and removed the irrelevant variables from a causal perspective. ADS [28] generated pseudo modality through adversarial training and mutual information regularization. Dual-Norm [34] assisted in selecting the optimal normalization path for unseen domain images through style information. SLAug [20] obtained augmented images by performing different intensity transformations on different regions of the mask. CCSDG [8] acquired style representation and structural representation through feature decoupling and only performed segmentation based on structural representation.

Qualitative comparison of our method with the competing methods is summarized in Table 1 and Table 2, showing the Dice coefficient results with BinRushed and Magrabia as the source domains, respectively. We also visualize prediction results of images from different domains in Fig. 3 and Fig. 4.

As shown in the Table 1 and Table 2, all the methods outperform the baseline due to the domain shift between the source domain used for training and the target domain used for testing, making it difficult for model trained simply using the source domain to generalize to unseen domains.

Among all the methods, our method achieves the highest average Dice coefficient. Compared with the methods designed for grayscale medical images, our method has a significant improvement, indicating that our method has more advantages in color medical images. This is mainly because the methods designed for grayscale images usually transform pixel intensity values in a single channel, while color images usually have three channels. The difference in the number of channels in color images can affect the effectiveness of these methods. Even if value transformation is performed separately for each channel of a color image, the effect is limited and cannot simulate complex variations.

In contrast, our augmentation method is based on the Fourier transform, which is the most commonly used method in multi-source setting to augment color medical images, and its effectiveness has been demonstrated in many works. Therefore, the effectiveness can be guaranteed when applied to single-source setting. At the same time, feature sensitivity can explicitly guide the network to learn domain-invariant features. The combination of them can significantly improve the generalization ability of the model.

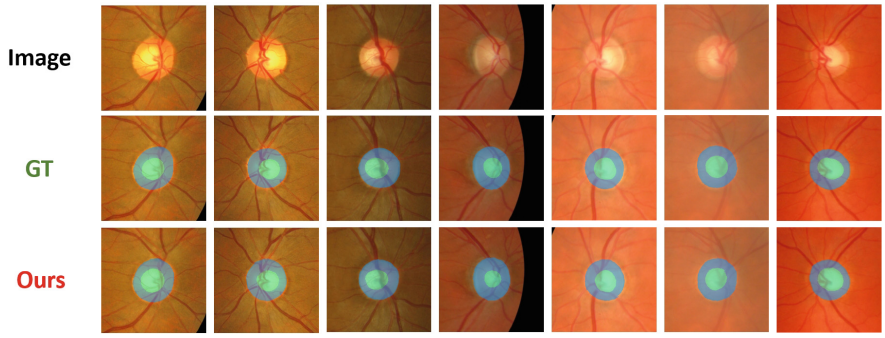


Fig. 3. Our method's visualization results of images from different domains

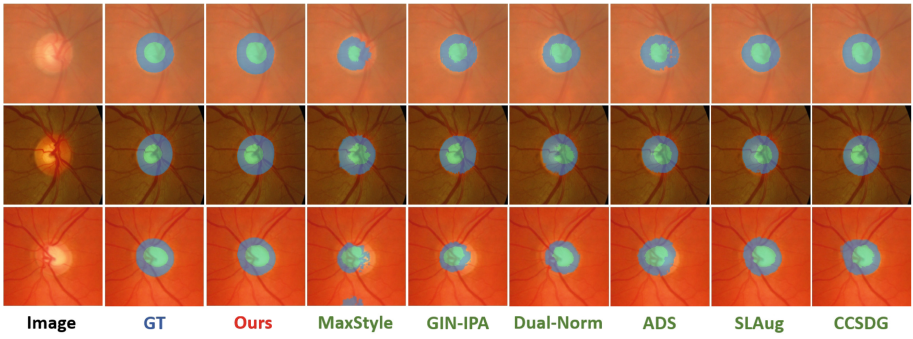


Fig. 4. Comparative visualization results of images from different domains predicted by our and other competing methods

The method CCSDG also learns domain-invariant features from the perspective of feature channels and is also the optimal method on this dataset, but our method is better than it, and the key reason is that our augmentation method makes the data more diverse, making the network can better learn domain-invariant features.

In Fig. 3, we visualize our results on images from different unknown domains, and Fig. 4 shows the comparison of our method's segmentation results with other methods. As can be seen from the figures, our method can more accurately segment target structures when encountering images from unseen domains and the

boundaries of segmentation results is relative smooth. Our method is able to maintain segmentation accuracy even when the domain shift in the unseen domains is significant, which is difficult for other methods to do so.

4.4 Ablation Analysis

Table 3. Ablation Experiments

U-net	SAPM	NormAug	Sensitivity	BinRushed		Magrabia	
				OD	*OC*	*OD*	*OC*
✓				85.37	73.46	86.59	76.88
✓	✓			94.98	84.41	94.55	84.05
✓		✓		90.67	80.12	90.13	79.64
✓		✓	✓	89.07	78.62	89.95	79.45
✓	✓		✓	**95.96**	**86.22**	**95.65**	**86.19**

To evaluate the effectiveness of SAPM and sensitivity guidance, we conduct ablation experiments using BinRushed and Magrabia as the source domain, respectively. Results of the ablation experiments are shown in Table 3. The first line is the result of the baseline U-net, the second line is the result of only using the SAPM proposed in this paper based on U-net, the third line is the result of only using common augmentation methods on u-net, such as color jittering and Gaussian blurring, the fourth line is the result obtained by replacing the SAPM with the common augmentation methods and using sensitivity to guide the network, and the last line is the result of the overall method of this article.

As we can see from Table 3, the results are greatly improved with the use of SAPM, which demonstrates that our augmentation method can greatly expand the diversity of data. Although normal augmentation methods can also compensate for the lack of cross-domain data to a certain extent, their effectiveness is far inferior to our proposed method. After further adopting sensitivity guidance, the performance of our network has been further improved.

However, after using sensitivity guidance on normal augmentation methods, the results do not show significant changes and even decreased. This may be due to the limited data diversity expansion ability of normal augmentation methods, which results in the guidance network being unable to effectively learn domain-invariant features and overfitting on the source domain. The results in the table also show that expanding data diversity is the most direct and effective way to improve the model's generalization ability.

We also conduct ablation experiments on the hyperparameters α and β. We use BinRushed as source domain and Base1, Base2, and Base3 as the target domains. The average results are shown in Fig. 5. We first fix α to 6 and vary β, and find that the generalization performance firstly increases and reaches a

maximum at β is 0.5, then the performance is stable or decreases slightly. When α takes other values, the overall influence of variation in β on performance is the same. Next, we fix β to 0.5 and vary α, the performance increases first and reaches its maximum when α is 6, and then the performance gradually decreases. Intuitively, β offers the baseline value to perturbation and thus has less impact on generalization performance after exceeding a certain threshold, while α controls overall intensity applied to perturbations, so that variation of α has a greater impact on the perturbation effect and performance. We set beta to 0.5 and alpha to 6 in this paper, but the optimal values may vary slightly depending on datasets and tasks.

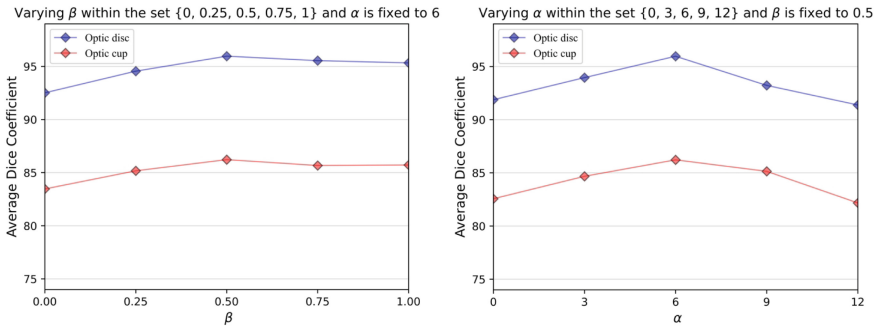


Fig. 5. Effect of hyperparameters α and β. α is fixed to 6 when varying β, β is fixed to 0.5 when varying α.

5 Conclusion

In this paper, we propose a novel Fourier-based data augmentation method called SAPM for color medical images, which can greatly expand the diversity of data without using information from other domains. The SAPM uses different Gaussian distributions to perturb different regions of the amplitude map, thereby avoiding the need for information from other domains. Also, we apply feature sensitivity guidance to guide the network to learn domain-invariant features to improve the generalization performance of the model further. The feature sensitivity guidance can guide the network to suppress feature channels sensitive to domain shift and emphasize feature channels insensitive to domain shift. Our results on the cross-domain fundus dataset demonstrates the effectiveness of SAPM and sensitivity guidance. The results of our ablation experiments show that SAPM can significantly improve the generalization performance of the network by expanding the data diversity. Meanwhile, the method proposed in this article also has certain limitations. For example, when encountering grayscale medical images, this method is not as effective as other methods specifically designed for grayscale images. We will continue explore and try to address these limitations in our future works.

Acknowledgements. This work was supported by the National Key Research and Development Program of China under Grant 2023YFB2904000 and the Natural Science Foundation of the Fujian Province, China, under Grant 2022J01574.

References

1. Almazroa, A., et al.: Retinal fundus images for glaucoma analysis: the RIGA dataset. In: Medical Imaging 2018: Imaging Informatics for Healthcare, Research, and Applications, vol. 10579, pp. 55–62. SPIE (2018)
2. Chen, C., Li, Z., Ouyang, C., Sinclair, M., Bai, W., Rueckert, D.: MaxStyle: adversarial style composition for robust medical image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 151–161. Springer, Cham (2022)
3. Choi, S., Jung, S., Yun, H., Kim, J.T., Kim, S., Choo, J.: RobustNet: improving domain generalization in urban-scene segmentation via instance selective whitening. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 11580–11590 (2021)
4. Decencière, E., et al.: Feedback on a publicly distributed image database: the messidor database. Image Anal. Stereol. **33**(3), 231–234 (2014)
5. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778 (2016)
6. 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)
7. Hu, S., Liao, Z., Xia, Y.: Domain specific convolution and high frequency reconstruction based unsupervised domain adaptation for medical image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 650–659. Springer (2022)
8. Hu, S., Liao, Z., Xia, Y.: Devil is in channels: contrastive single domain generalization for medical image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 14–23. Springer (2023)
9. Huang, Z., Wang, H., Xing, E.P., Huang, D.: Self-challenging improves cross-domain generalization. In: ECCV 2020, Part II, pp. 124–140. Springer (2020)
10. Isensee, F., Jaeger, P.F., Kohl, S.A., Petersen, J., Maier-Hein, K.H.: nnU-net: a self-configuring method for deep learning-based biomedical image segmentation. Nat. Methods **18**(2), 203–211 (2021)
11. Li, H., et al.: Frequency-mixed single-source domain generalization for medical image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 127–136. Springer (2023)
12. Li, L., et al.: Progressive domain expansion network for single domain generalization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 224–233 (2021)
13. Litjens, G., et al.: A survey on deep learning in medical image analysis. Med. Image Anal. **42**, 60–88 (2017)
14. Ouyang, C., et al.: Causality-inspired single-source domain generalization for medical image segmentation. IEEE Trans. Med. Imaging **42**(4), 1095–1106 (2022)
15. Pan, H., et al.: Domain generalization with Fourier transform and soft thresholding. arXiv preprint arXiv:2309.09866 (2023)

16. Qiao, F., Zhao, L., Peng, X.: Learning to learn single domain generalization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 12556–12565 (2020)
17. Ronneberger, O., Fischer, P., Brox, T.: U-net: convolutional networks for biomedical image segmentation. In: MICCAI 2015, Part III, pp. 234–241. Springer (2015)
18. Seo, S., Suh, Y., Kim, D., Kim, G., Han, J., Han, B.: Learning to optimize domain specific normalization for domain generalization. In: ECCV 2020, Part XXII, pp. 68–83. Springer (2020)
19. Shen, D., Wu, G., Suk, H.I.: Deep learning in medical image analysis. Annu. Rev. Biomed. Eng. **19**, 221–248 (2017)
20. Su, Z., Yao, K., Yang, X., Huang, K., Wang, Q., Sun, J.: Rethinking data augmentation for single-source domain generalization in medical image segmentation. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 37, pp. 2366–2374 (2023)
21. Ulyanov, D., Vedaldi, A., Lempitsky, V.: Instance normalization: the missing ingredient for fast stylization. arXiv preprint arXiv:1607.08022 (2016)
22. Wang, J., et al.: Generalizing to unseen domains: a survey on domain generalization. IEEE Trans. Knowl. Data Eng. (2022)
23. Wang, S., Yu, L., Li, K., Yang, X., Fu, C.W., Heng, P.A.: DoFE: domain-oriented feature embedding for generalizable fundus image segmentation on unseen datasets. IEEE Trans. Med. Imaging **39**(12), 4237–4248 (2020)
24. Wang, Z., Luo, Y., Qiu, R., Huang, Z., Baktashmotlagh, M.: Learning to diversify for single domain generalization. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 834–843 (2021)
25. Wilson, G., Cook, D.J.: A survey of unsupervised deep domain adaptation. ACM Trans. Intell. Syst. Technol. (TIST) **11**(5), 1–46 (2020)
26. Xu, Q., et al.: DIRL: domain-invariant representation learning for generalizable semantic segmentation. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 36, pp. 2884–2892 (2022)
27. Xu, Q., Zhang, R., Zhang, Y., Wang, Y., Tian, Q.: A Fourier-based framework for domain generalization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 14383–14392 (2021)
28. Xu, Y., Xie, S., Reynolds, M., Ragoza, M., Gong, M., Batmanghelich, K.: Adversarial consistency for single domain generalization in medical image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 671–681. Springer (2022)
29. Xu, Z., Liu, D., Yang, J., Raffel, C., Niethammer, M.: Robust and generalizable visual representation learning via random convolutions. arXiv preprint arXiv:2007.13003 (2020)
30. Yao, H., Hu, X., Li, X.: Enhancing pseudo label quality for semi-supervised domain-generalized medical image segmentation. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 36, pp. 3099–3107 (2022)
31. Zhou, K., Liu, Z., Qiao, Y., Xiang, T., Loy, C.C.: Domain generalization: a survey. IEEE Trans. Pattern Anal. Mach. Intell. **45**(4), 4396–4415 (2022)
32. Zhou, K., Yang, Y., Qiao, Y., Xiang, T.: Domain generalization with mixstyle. In: International Conference on Learning Representations (ICLR) (2021)
33. Zhou, Z., Qi, L., Shi, Y.: Generalizable medical image segmentation via random amplitude mixup and domain-specific image restoration. In: European Conference on Computer Vision, pp. 420–436. Springer (2022)

34. Zhou, Z., Qi, L., Yang, X., Ni, D., Shi, Y.: Generalizable cross-modality medical image segmentation via style augmentation and dual normalization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 20856–20865 (2022)
35. Zhou, Zet al.: Models genesis: generic autodidactic models for 3D medical image analysis. In: MICCAI 2019, Part IV, pp. 384–393. Springer (2019)

Transformer Models for Enhanced Calcifications Detection in Mammography

Marco Cantone[1(✉)], Claudio Marrocco[1], Francesco Tortorella[2], and Alessandro Bria[1]

[1] Department of Electrical and Information Engineering, University of Cassino and Southern Latium, Via Gaetano Di Biasio 43, 03043 Cassino, FR, Italy
{marco.cantone,c.marrocco,a.bria}@unicas.it

[2] Department of Information and Electrical Engineering and Applied Mathematics, University of Salerno, 84084 Fisciano, SA, Italy
ftortorella@unisa.it

Abstract. In recent years, the image analysis landscape is witnessing a paradigm shift with the emergence of the vision transformer as a better alternative to Convolutional Neural Networks (CNNs). Transformers process sequences globally with self-attention capturing long-range features, while CNNs extract features locally through convolutional operations. We propose the adoption of Swin Transformer as backbone for calcification cluster detection in mammography, assessing its efficacy through a comprehensive experimental study comparing transformer-based and CNN-based models. Our experiments conducted on the large-scale mammography image database OMI-DB demonstrate a notable superiority of the Swin Transformer architecture. The best-performing Swin backbone obtained a sensitivity of 80.67% at 0.1 false positive per image, with a +3.34% improvement over the best convolutional backbone. Our findings underscore the efficacy of transformer-based architectures for detecting clusters of calcifications in mammography, offering improved diagnostic accuracy in this field.

Keywords: Calcifications detection · Mammography · Transformers

1 Introduction

Breast cancer is the most common cancer among women and is the second leading cause of death. Throughout the years, the incidence of breast cancer has risen worldwide, and one million new cases are reported annually [13,19]. The early diagnosis of breast cancer is essential for improving survival chances, underscoring the adoption of screening programs in many countries. Mammography stands out as one of the most widely utilized imaging modalities both for screening and diagnostic purposes [17]. Calcifications are tiny deposits of calcium that appear as bright spots in mammography images, and they are recognized among the initial discernible indicators of breast cancer. Moreover, a multitude of breast lesions exhibit associations with calcifications. [2,35].

A. Antonacopoulos et al. (Eds.): ICPR 2024, LNCS 15313, pp. 17–33, 2025.
https://doi.org/10.1007/978-3-031-78201-5_2

Computer Aided Diagnosis (CAD) systems are designed to help radiologists in analyzing medical images. Traditionally, these systems are based on image processing techniques and hand-crafted features designed by human experts. Early approaches to the detection of calcifications include Difference of Gaussian (DoG) filters, thresholding, and morphological operation [9,30]. With the advancement of Machine Learning (ML), many of these approaches have been used in conjunction with powerful statistical models and frequently employed for data preprocessing [14,29,37]. Recent improvements in medical image analysis using Deep Learning (DL) contribute to enhancing the performance of CAD systems [17,23]. In the last decade, the most popular approaches have been based on Convolutional Neural Networks (CNNs). These models are composed of convolutional layers, designed to automatically learn spatial hierarchies of features from raw input data. CNNs have been widely applied for mammography analysis, including single calcification detection and cluster detection [1,6,12,15,31,33].

In 2017, the introduction of transformer architecture [36] established new state-of-the-art (SOTA) in many different fields, even in the medical image domain. Although the transformer has replaced many specialized neural architectures for several domains, its superiority remains uncertain across all scenarios, considering its high demand for training data and the lack of certain biases, such as locality [5]. However, the sparse and scattered nature of calcifications lends itself well to the global contextual understanding provided by transformers and the capability of attention to correlate various parts of the image. Furthermore, the small patch size adopted by some recent visual transformers, like the Swin Transformer [25] with a patch size of 4, can be particularly effective in capturing the subtle variations in the image associated with small lesions such as calcifications. Transformer-based models have been applied for various tasks related to mammography analysis including single-view and multi-view mammography classification [5,8], mass segmentation [24] and mass detection [3].

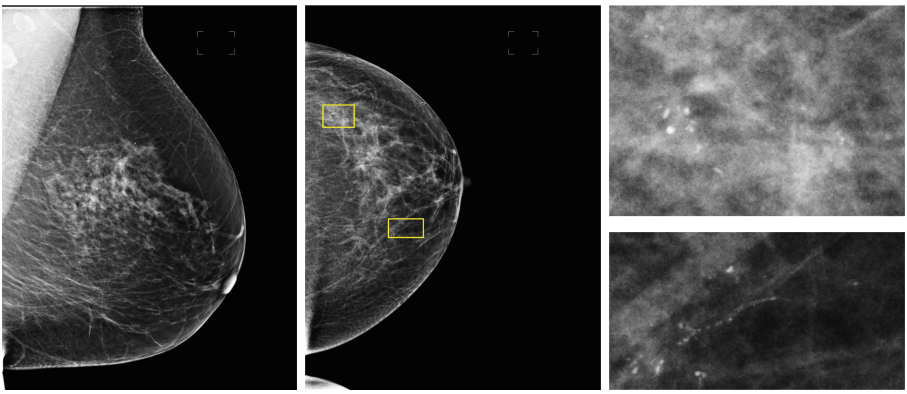
The main contributions of this work are twofold. We propose the adoption of the Swin Transformer, a hierarchical vision transformer backbone, as multi-scale feature extractor for the detection of calcification clusters in mammography images. Moreover, we investigate the benefits of using transformers through comprehensive experimentation using different convolutional backbone architectures in combination with three object detection heads on the OMI-DB dataset [16]. To the best of our knowledge, this is the first work applying transformer models to the detection of calcification clusters.

The rest of this work is organized as follows. Section 2 describes the dataset and the network models employed. Details about the experimental methodology, the implementation, and the metrics used are provided in Sect. 3. In Sect. 4 the obtained results are presented and discussed. We conclude with a summary and a critical discussion in Sect. 5.

2 Materials

2.1 Datasets

OMI-DB [16] is a large mammography database, the creation of which was funded by Cancer Research UK. The dataset contains images in DICOM format coupled with anonymised clinical information, including bounding boxes and lesion type annotations. The images include both *for processing* and *for presentation* mammograms from scanners of different vendors such as Hologic Inc., Siemens, Philips, General Electric Medical System, and Bioptics Inc. For this study, only *for presentation* images from Hologic Inc. scanners were selected as they represented the vast majority of the dataset. For training DL models we used only the images suitable for calcification detection. Two types of images were selected: normal mammograms with no lesions present, and images associated with malignancies containing one or more calcification clusters. Images with calcification clusters resulting in a benign biopsy were discarded and not included either as normal or positive. Visual inspection of all the selected images was performed to obtain a clean dataset without unwanted objects such as implants, marker clips, bands across the image and overlaid text. The dataset obtained consists of $9,895$ normal images and $2,563$ mammograms containing $2,962$ calcification clusters. We saved all mammograms in 16-bit PNG format for faster processing with respect to DICOM format. In Fig. 1 are reported examples taken from the OMI-DB dataset.



(a) (b) (c)

Fig. 1. Example of mammograms from the OMI-DB dataset. (a) a normal image, (b) a malignant image with two clusters of calcifications, (c) magnification of the two clusters present in (b).

2.2 Backbones

ResNet. In 2015 He et al. proposed the ResNet architecture [18] for addressing the vanishing gradient problem encountered in very deep networks. They introduce skip connections which enable the flow of information from earlier layers to later layers by bypassing intermediate layers. This facilitates the training of deeper networks by allowing the gradients to propagate more effectively during backpropagation. In ResNet, each layer learns residual functions with reference to the layer inputs, rather than directly learning underlying mapping functions.

ResNetStrikesBack. Taking advantage of methodological innovations in neural network training strategies and data augmentation, Wightman et al. [38] trained a ResNet-50 with a procedure that integrates such advances. With the new training setting, a vanilla ResNet-50 managed to achieve an 80.4% top-1 accuracy on ImageNet [10] without extra data or distillation, a big improvement compared to the 75.3% obtained in the original work.

EfficientNet. Characterized by the efficient use of computational resources, EfficientNet [34] is a CNN that employs a compound scaling method that uniformly scales the networks depth, width, and resolution to balance model complexity and computational cost. The baseline model on which compound scaling is applied is obtained by leveraging a multi-objective neural architecture search that optimizes both accuracy and FLOPS. These design principles make EfficientNet well-suited for resource-constrained environments and applications where computational efficiency is critical.

Swin. The Shifted Window Transformer [25] is a transformer-based architecture that incorporates hierarchical processing of image patches to capture both local and global contextual information effectively. Unlike traditional convolutional neural networks, which process images in a sequential manner, Swin Transformer organizes image patches into a hierarchical structure and processes them through multiple stages, each consisting of alternating layers of local and global self-attention mechanisms. This hierarchical processing enables Swin Transformer to capture information at different scales efficiently, facilitating better modeling of spatial relationships within images. Moreover, Swin Transformer introduces shifted windows to capture long-range dependencies effectively while maintaining linear computational efficiency. Swin Transformer achieved SOTA performance in various computer vision tasks, including image classification, object detection, and semantic segmentation. In Table 1 are reported the architectural parameters for the Swin models employed in this work.

ConvNeXt. In 2022 Liu et al. present ConvNeXt [26], a modified variant of the ResNet-50 inspired by the architectural innovations of the Swin Transformer. Mimicking Swins macro design, ConvNeXt introduces changes regarding the

Table 1. Architectural parameters for the two Swin Transformers variants employed in this work.

	depths	num. heads	embedding dim.	dropout	drop path rate	windows size	patch size
Swin-T	[2, 2, 6, 2]	[3, 6, 12, 24]	96	0	0.2	7	4
Swin-B	[2, 2, 18, 2]	[4, 8, 16, 32]	128	0	0.3	7	4

number of layers in each block and embraces patch-based image representations. Furthermore, micro-level refinements such as grouped convolution and the adoption of GeLU activation functions are employed. Remarkably, ConvNeXt achieves competitive performance without resorting to self-attention, challenging attention mechanism as the main actor for achieving competitive performance.

2.3 Object Detection Heads

RetinaNet. Anchor boxes were introduced in the field of object detection with the RetinaNet [21] architecture. They are predefined bounding boxes of various sizes and aspect ratios, allowing the model to efficiently detect objects across different scales and orientations in images. The RetinaNet head consists of two key subnetworks: the classification subnet and the box regression subnet. The classification subnet employs a series of convolutional layers to generate class predictions for each anchor box. Meanwhile, the box regression subnet utilizes similar convolutional layers to predict bounding box displacement, refining the initial anchor box proposals. The focal loss function dynamically adjusts the loss contribution of each anchor box based on its classification difficulty. This loss mechanism effectively mitigates the impact of class imbalance, allowing RetinaNet to achieve superior performance on object detection tasks across various datasets and benchmarks.

RepPoints. Introduced by Jiang et al. in 2020 [39], the RepPoints head approaches the object detection task by leveraging representative points for precise localization and feature representation. RepPoints focuses on compact descriptors rather than bounding boxes or anchor points, enhancing adaptability to diverse object shapes and sizes. Its architecture comprises a regression subnet for refining object proposals and a representative point generation module for accurate localization.

DDETR. The Deformable Detection Transformer was proposed by Zhu et al. [40] by addressing the limitation of the DETR [7] regarding feature spatial resolution and convergence speed. It achieves this by combining the best of the sparse spatial sampling of deformable convolution, and the relation modeling capability of Transformers. It proposed the deformable attention module, which attends to a small set of sampling locations as a pre-filter for prominent

key elements out of all the feature map pixels. Multiscale deformable attention modules facilitate the effective handling of spatial information across different scales and enhance model robustness to object size variations.

3 Experimental Methodology

DL models for object detection comprise a backbone that extracts features from the raw input image and a network head that localizes and classifies the objects returning labels and bounding boxes as output. In this work, we propose the Swin Transformer as backbone for calcification cluster detection, comparing its efficacy against widely used CNNs through an extensive experimental study. Overall we used 8 backbone models: ResNet50, ResNet101, ResNetStrikesBack, EfficientNet, ConvNeXt-T, ConvNeXt-S, Swin-T, Swin-B, and 3 heads: RetinaNet, RepPoints and DDETR. We train and test each backbone-head combination resulting in 24 experiments. All the backbones were pretrained on ImageNet [10] whereas the different network heads were pretrained on COCO [22] then the entire architecture was fine-tuned on our dataset.

3.1 Data Preprocessing

The following data preprocessing was applied. First, we segmented the breast area discarding as much background as possible. This reduced the image size speeding up the training and allowing higher resolution and batch size. Then, pixel values were normalized to zero mean and unit standard deviation and the images were resized to 1280×800 resolution. In order to use the model weights pretrained on ImageNet and COCO, we convert all the images to RGB by replicating the grayscale channel.

3.2 Data Augmentation

Following the work of Betancourt Tarifa et al. [3] on the mass detection in mammography, we applied the following data augmentation techniques, each one with a probability of 50%: (i) horizontal flip; (ii) random crop; (iii) contrast transformation, with magnitude values of [0.4, 0.8, 1.5]; and (iv) brightness transformation, with magnitude values of [0.3, 0.7, 1.3]. For Swin-B and ResNet101 backbones the probabilities were increased to 60%.

3.3 Training Hyperparameters

The dataset was split randomly using a 70-10-20 ratio in train, validation, and test set. We trained the models for a maximum number of epochs ranging from 30 to 100. The best model was selected by mean Average Precision (mAP) over IoU thresholds from 0.1 to 0.5 with a step of 0.05. We employed either Stochastic Gradient Descent or AdamW [27] using different learning rates and a batch size of 2. We adopt an exponential decay learning rate scheduler with linear

warmup with different rates of decay and step epoch. In Table 2 are reported the hyperparameters for all the trained architectures. Optimizations were conducted exclusively on the validation set.

Table 2. Training hyperparameters. γ indicates the learning rate decay and the step column refers to epochs after which the learning rate is adjusted.

Backbone	Head	Optimizer	LR	Best epoch (total)	γ	Step
ResNet50	RetinaNet	SGD	7.81×10^{-5}	17 (30)	0.2	[6, 12, 18, 24]
	RepPoints	SGD	1.00×10^{-4}	12 (30)	0.1	[6, 12, 18, 24]
	DDETR	AdamW	1.25×10^{-5}	17 (50)	0.1	[40]
ResNet101	RetinaNet	SGD	7.81×10^{-5}	13 (30)	0.2	[6, 12, 18, 24]
	RepPoints	SGD	1.00×10^{-4}	16 (30)	0.2	[6, 12, 18, 24]
	DDETR	AdamW	1.25×10^{-5}	23 (50)	0.1	[40]
ResNet- StrikesBack	RetinaNet	SGD	1.00×10^{-4}	27 (50)	0.1	[6, 12, 18, 24]
	RepPoints	AdamW	1.25×10^{-5}	21 (40)	0.1	[36]
	DDETR	AdamW	1.25×10^{-5}	30 (100)	0.1	[40]
EfficientNet	RetinaNet	SGD	1.00×10^{-4}	15 (30)	0.1	[6, 12, 18, 24]
	RepPoints	AdamW	1.00×10^{-4}	15 (30)	0.4	[6, 12, 18, 24]
	DDETR	AdamW	1.25×10^{-5}	65 (100)	0.1	[40]
ConvNeXt-T	RetinaNet	AdamW	1.25×10^{-5}	85 (100)	0.1	[36, 44]
	RepPoints	AdamW	1.25×10^{-5}	60 (100)	0.1	[36, 44]
	DDETR	AdamW	1.25×10^{-5}	29 (100)	0.1	[40]
ConvNeXt-S	RetinaNet	AdamW	1.25×10^{-5}	96 (100)	0.1	[36, 44]
	RepPoints	AdamW	1.25×10^{-5}	74 (100)	0.1	[30, 45, 60]
	DDETR	AdamW	1.25×10^{-5}	35 (100)	0.1	[40]
Swin-T	RetinaNet	AdamW	1.25×10^{-5}	12 (30)	–	–
	RepPoints	AdamW	1.25×10^{-5}	30 (50)	0.1	[36, 44]
	DDETR	AdamW	1.25×10^{-5}	35 (50)	0.1	[40]
Swin-B	RetinaNet	AdamW	1.25×10^{-5}	19 (30)	–	–
	RepPoints	AdamW	1.25×10^{-5}	20 (30)	–	–
	DDETR	AdamW	1.25×10^{-5}	18 (50)	0.1	[40]

3.4 Performance Evaluation

To evaluate the performances of the employed architectures, we calculated cluster-based Free Receiver Operating Characteristic (FROC) curves that report the True Positive Rate (TPR) over the average number of False Positives per Image (FPpI) by varying the decision threshold applied to the scores associated with the detected object. A predicted box was considered a true positive when its IoU with the groundtruth cluster bounding box was equal or greater than 0.1. All predictions on normal images were counted as false positives. From the FROC curve we extract 3 metrics: the Area Under the FROC Curve (AUFC) in the FPpI ranges [0, 0.1] and [0, 1], and the TPR at 0.1 FPpI.

3.5 Statistical Analysis

To assess the statistical relevance of differences in performance metrics between pairs of backbones sharing the same head, the bootstrap method [32] was applied. We sampled patients with replacement 10,000 times, with each bootstrap sample containing the same number of patients as the original set. At each bootstrapping iteration, FROC curves were recalculated for each method, and differences in the metrics considered between methods under comparison were evaluated. p-values were computed as the fraction of performance differences that were negative or zero, corresponding to cases where the target method did not outperform the method compared (null hypothesis). Performance differences were considered statistically significant if p-value < 0.05.

4 Results and Discussion

In Tables 3 and 4 are reported the values of TPR at 0.1 FPpI, $AUFC$ in the ranges [0, 0.1] and [0, 1] for each backbone-head combination tested. Across all the heads and metrics considered, the Swin-B backbone demonstrates superior performance compared to other backbones employed, achieving an average +4.11% TPR with respect to the top-performing convolutional backbone, ConvNeXt-S. Swin-T, a less complex variant of Swin-B, did not surpass the convolutional counterpart across all heads and metrics. However, on average it performs better than other CNNs, exhibiting a TPR increment of +2.06% compared to ConvNeXt-S, despite employing fewer parameters. Among the heads, the best result was yelded by RepPoints, followed by RetinaNet and DDETR with a TPR of 80.67%, 80.00%, and 78.67%, respectively. The absolute highest result was achieved by RepPoints/Swin-B with a 80.67% TPR, a 71.13% $AUFC_{[0,0.1]}$ and a 86.83% $AUFC_{[0,1]}$. In Fig. 2 and Fig. 3 models outputs and radiologist annotation are represented on examples images taken from the test set. The results indicate Swin-B as an effective alternative over CNN backbone for cluster detection in mammography. This can be due to a more effective features extraction since the results are the best across all the heads employed. Additionally, the RepPoints/Swin-B architecture, which features a combination of a transformer backbone and a convolutional head, highlights the importance of integrating these two different paradigms.

Figure 4 shows the FROC curves for all the models employed with a statistical comparison between the best transformer and convolutional backbone for each head, selected by $AUFC_{[0,0.1]}$. For the RetinaNet and RepPoints heads the FROC of the Swin-B is always higher then all the others, except for a small range near 0.01 FPpI where it is surpassed by the EfficientNet in the case of RetinaNet head, and ConvNeXt-S in the case of RepPoints. For the DDETR head, the Swin-B and the ConvNeXt-S achieve comparable performance. The bottom-right plot of Fig. 4 shows a clear overlap between the two FROCs. In general, the sensitivity between Swin-B and the best convolution backbone is comparable in

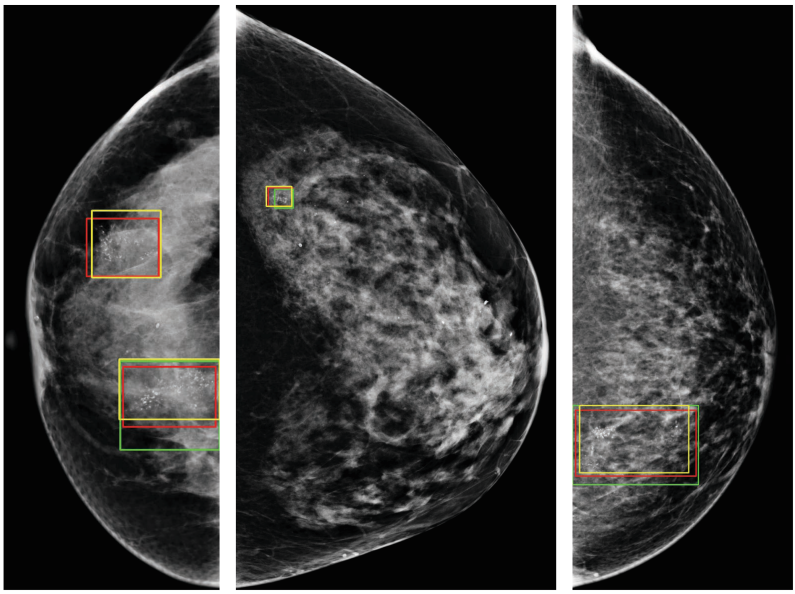
Table 3. TPR at 0.1 FPpI for each backbone-head combination. In bold the best result obtained for each head.

	TPR at 0.1 FPpI		
	RetinaNet	RepPoints	DDETR
ResNet50	70.17%	71.83%	67.67%
ResNet101	71.00%	74.33%	70.17%
ResNetStrikesBack	74.17%	74.50%	69.83%
EfficientNet	75.00%	77.33%	63.67%
ConvNeXt-T	70.00%	56.33%	76.50%
ConvNeXt-S	76.50%	74.00%	76.50%
Swin-T	77.17%	79.17%	76.83%
Swin-B	**80.00%**	**80.67%**	**78.67%**

Table 4. AUFC for each backbone-head combination. In bold the best result obtained for each head.

	$AUFC_{[0,0.1]}$			$AUFC_{[0,1]}$		
	RetinaNet	RepPoints	DDETR	RetinaNet	RepPoints	DDETR
ResNet50	56.27%	61.39%	55.12%	80.46%	81.78%	76.67%
ResNet101	59.69%	63.33%	57.73%	81.64%	82.39%	79.99%
ResNetStrikesBack	63.40%	63.99%	54.98%	83.18%	83.00%	79.59%
EfficientNet	64.37%	65.17%	48.73%	83.45%	84.44%	73.80%
ConvNeXt-T	58.07%	45.11%	63.81%	77.29%	65.60%	83.12%
ConvNeXt-S	64.15%	66.14%	61.84%	81.09%	78.20%	84.09%
Swin-T	63.41%	70.20%	63.98%	86.03%	86.26%	83.50%
Swin-B	**68.36%**	**71.13%**	**65.60%**	**86.27%**	**86.83%**	**84.39%**

the FPpI range [0.01, 0.03] while after these values the transformer-based backbone clearly surpasses all the convolutional models. Swin-B consistently outperforms convolutional backbones, particularly in higher false positive rate ranges, indicating its robustness in handling challenging detection scenarios. We believe that the Swin Transformers hierarchical representation learning and spatial context awareness contributed to its superior performance for calcification cluster detection in mammography. By employing a self-attention mechanism, the model captures intricate patterns at multiple scales, effectively discerning calcification clusters from surrounding breast tissue. This hierarchical approach allows the Swin Transformer to encode complex spatial relationships within mammogram images, enabling it to effectively differentiate between true calcification clusters and background noise or artifacts. The models ability to integrate spatial context information across the entire image facilitates robust detection by considering the relative positions and interactions between pixels and regions.



(a) RetinaNet (b) RepPoints (c) DDETR

Fig. 2. Example images from the test set with overlaid annotations and network bounding boxes, with each subplot referring to a different head. In red the models outputs using the Swin-B as backbone; in green, the models outputs using the best convolutional backbone for the specific head selected by maximizing the $AUFC_{[0,0.1]}$; in yellow the radiologist annotation. (Color figure online)

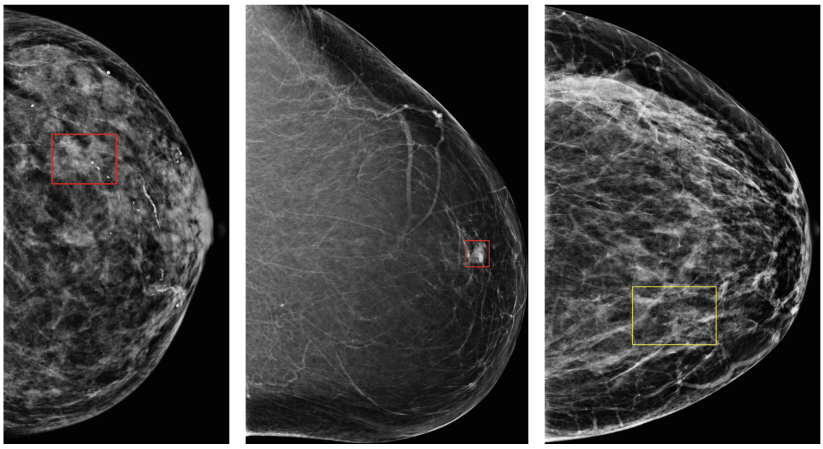


Fig. 3. Example images from the test set with overlaid annotations (yellow) and RepPoints/Swin-B predicted bounding boxes (red) at 0.5 threshold score. (a) and (b) show false positive examples, and (c) an undetected cluster. (Color figure online)

Table 5 illustrates the computational demand in GFLOPs for each detector model.

Table 5. GFLOPs for each backbone-head combination.

	ResNet50	ResNet101	ResNetStrikesBack	EfficientNet	ConvNeXt-T	ConvNeXt-S	Swin-T	Swin-B
RetinaNet	206	282	204	117	562	648	211	444
RepPoints	190	266	190	102	498	584	195	428
DDETR	195	271	195	108	564	651	516	749

In Table 6 a comparison with existing methods is reported. The results are not directly comparable since they were obtained with different datasets and at different FPpIs. It can be observed that the proposed approach yields significantly lower false-positive values compared to those typically reported in the literature while maintaining a high TPR.

Table 6. Comparison with SOTA methods for calcification clusters detection.

	Dataset	TPR	FPpI
Gallardo et al., 2012 [11]	DDSM	0.82	2.55
Bria et al., 2016 [4]	Private dataset	0.96	0.21
Karale et al., 2019 [20]	InBreast	1	1.78
Rehman et al., 2021 [31]	DDSM	0.97	2.35
Cantone et al., 2023 [6]	OMI-DB	0.44	0.1
Ours	OMI-DB	0.81	0.1

4.1 Statistical Analysis

In Table 7 a statistical comparison between Swin-B and the best convolutional backbone for each head is reported. For the RetinaNet and RepPoints heads, the superiority of Swin-B was statistically relevant with a p-value always less than 0.018, and a TPR increment of +5.2 against the RetinaNet head, and +6.8 against the RepPoints head. The DDETR/Swin-B was not statistically better than DDETR/ConvNeXt-S obtaining p-values sligtly greater than 0.05 for all the metrics considered. However, the DDETR was the worst-performing head among the three tested, indicating that is not best suited for this task. This supports the idea that transformers are not always the best choice since DDETR, a transformer-based head, performs worst compared to the two convolution heads RetinaNet and RepPoints.

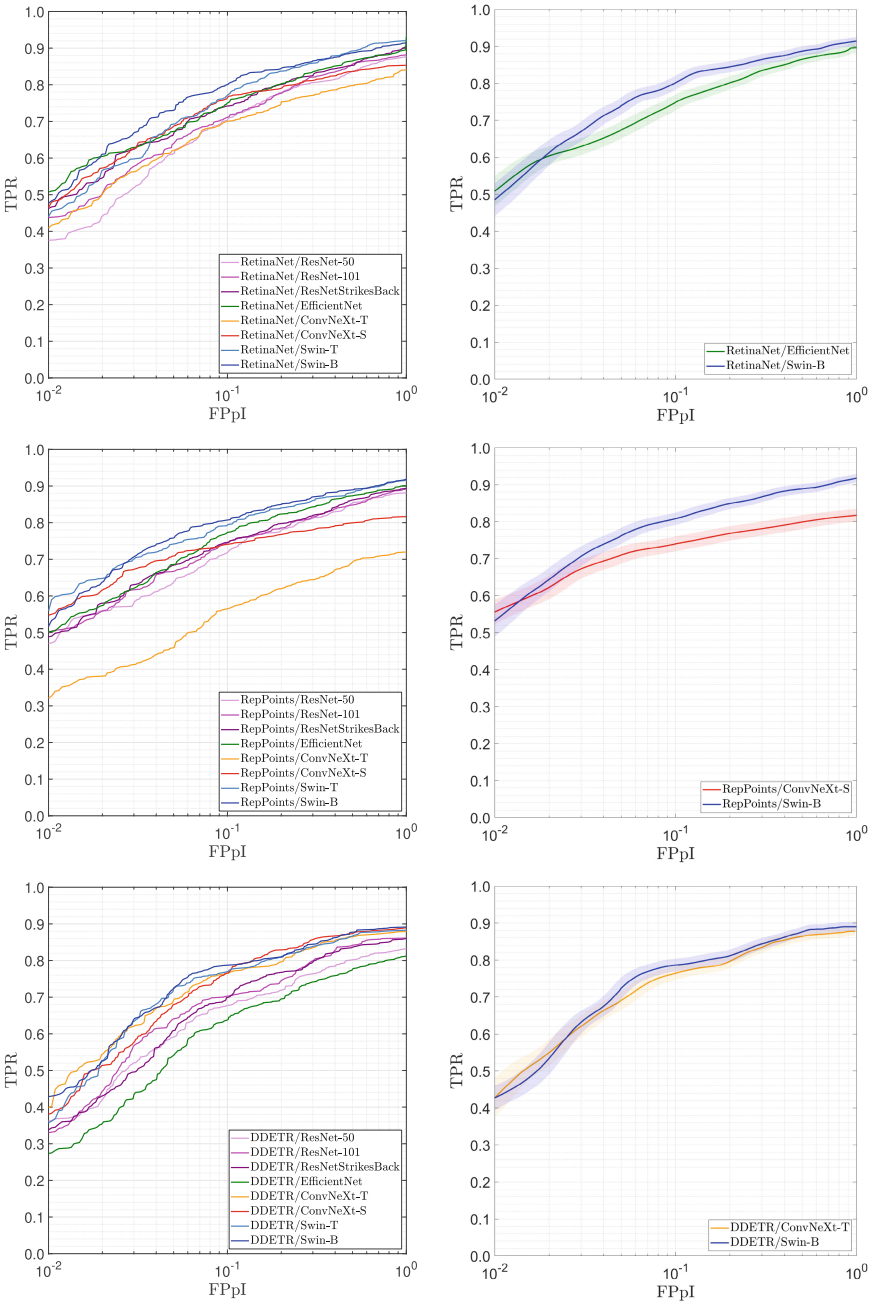


Fig. 4. Left: FROC curves illustrating the performance comparison of all tested backbones, with each subplot representing a different head. Right: Average FROC curves obtained from 10,000 bootstrap iterations illustrating the comparison between Swin-B and the best-performing CNN backbone for each head.

Table 7. Statistical comparison between Swin-B and best convolutional backbone selected by $AUFC_{[0,0.1]}$ using bootstrap method with 10,000 resampling.

Head	Backbone	ΔTPR (p-value)	$\Delta AUFC_{[0,0.1]}$ (p-value)	$\Delta AUFC_{[0,1]}$ (p-value)
RetinaNet	Swin-B vs. EfficientNet	+5.2 (0.0035)	+4.0 (0.0183)	+4.0 (0.0012)
RepPoints	Swin-B vs. ConvNeXt-S	+6.8 (<0.0001)	+5.0 (0.0022)	+8.6 (<0.0001)
DDETR	Swin-B vs. ConvNeXt-T	+2.1 (0.0947)	+1.8 (0.1660)	+1.3 (0.1038)

4.2 External Dataset Evaluation

In this section we evaluate the detection performance of RepPoints/Swin-B and RepPoints/ConvNeXt-S on the InBreast [28] dataset without retraining the models. The dataset consists of 105 normal images and 21 positive images with 27 annotated clusters. Figure 5 and Table 8 illustrate the obtained results. Also on the InBreast dataset, the Swin Transformer statistically significantly outperforms its convolutional counterpart, achieving an increase of 20.8% in TPR, 14.5% in $AUFC_{[0,0.1]}$, and 4.0% in $AUFC_{[0,1]}$. Moreover, the performance is superior to that achieved on the OMI-DB dataset, indicating a strong generalization capability.

Table 8. Comparison between RepPoints/Swin-B and RepPoints/ConvNeXt-S on InBreast without fine-tuning. The statistical analysis was carried out using bootstrap method with 10,000 resampling.

Metric	RepPoints/Swin-B	RepPoints/ConvNeXt-S	Δ-metric	p-value
TPR@0.1FPpI	95.6%	74.8%	+20.8%	0.0150
$AUFC_{[0,0.1]}$	71.8%	57.3%	+14.5%	0.0392
$AUFC_{[0,1]}$	87.1%	83.1%	+4.0%	0.0484

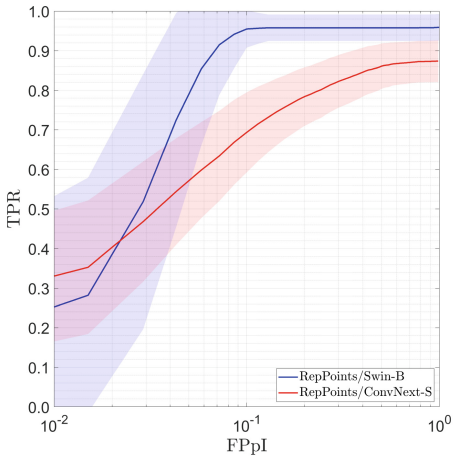


Fig. 5. Average FROC curves obtained from 10,000 bootstrap iterations illustrating the comparison between RepPoints/Swin-B and RepPoints/ConvNeXt-S on the InBreast dataset without fine-tuning.

5 Conclusions

In this work, we adopted the Swin Transformer as backbone for calcifications clusters detection in mammography comparing its performance with different CNNs through a comprehensive experimental study. The hierarchical long-ranges features extracted by the Swin Transformer consistently yielded superior performances across all heads, indicating the extraction of more valuable features. The best model achieved a remarkable result of 80.67% TPR at 0.1 FPpI and 86.83% $AUFC$ in the range $[0, 1]$, largely surpassing the best convolutional model RepPoints/ConvNeXt-S by +6.8 TPR and +8.6 $AUFC_{[0,1]}$ with high statistical significance (p-value < 0.0001). Relying exclusively on transformer-based models may not always yield optimal results, and combining elements from transformer and convolutional networks, as exemplified by the RepPoints/Swin-B model in our study, leads to superior performance for the detection of clusters of calcifications. These insights underscore the potentiality of transformer-based architectures as backbone networks for detecting sparse lesions in medical imaging.

Acknowledgements. This research was funded by Italian Ministry of University, MIUR program "Department of Excellence" Law 232/216 and by D.M. 351/2022 "Innovative PhD Programs for Public Administration". The mammography images and data used in this research were derived from the OPTIMAM imaging database (OMI-DB) [16]. We would like to acknowledge the OPTIMAM project team and staff at the Royal Surrey NHS Foundation Trust who developed the OPTIMAM database, Cancer Research UK which funded the creation and maintenance of OPTIMAM database and Cancer Research Horizons which facilitates access to the OPTIMAM data.

References

1. Abdelrahman, L., Al Ghamdi, M., Collado-Mesa, F., Abdel-Mottaleb, M.: Convolutional neural networks for breast cancer detection in mammography: a survey. Comput. Biol. Med. **131**, 104248 (2021)
2. Azam, S., et al.: Mammographic microcalcifications and risk of breast cancer. Br. J. Cancer **125**(5), 759–765 (2021)
3. Betancourt Tarifa, A.S., Marrocco, C., Molinara, M., Tortorella, F., Bria, A.: Transformer-based mass detection in digital mammograms. J. Ambient. Intell. Humaniz. Comput. **14**(3), 2723–2737 (2023)
4. Bria, A., Marrocco, C., Karssemeijer, N., Molinara, M., Tortorella, F.: Deep cascade classifiers to detect clusters of microcalcifications. In: IWDM, pp. 415–422. Springer (2016)
5. Cantone, M., Marrocco, C., Tortorella, F., Bria, A.: Convolutional networks and transformers for mammography classification: an experimental study. Sensors **23**(3), 1229 (2023)
6. Cantone, M., Marrocco, C., Tortorella, F., Bria, A.: Learnable dog convolutional filters for microcalcification detection. Artif. Intell. Med. **143**, 102629 (2023)
7. Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., Zagoruyko, S.: End-to-end object detection with transformers. In: European Conference on Computer Vision, pp. 213–229. Springer (2020)
8. Chen, X., et al.: Transformers improve breast cancer diagnosis from unregistered multi-view mammograms. Diagnostics **12**(7), 1549 (2022)
9. Cheng, H.D., Cai, X., Chen, X., Hu, L., Lou, X.: Computer-aided detection and classification of microcalcifications in mammograms: a survey. Pattern Recogn. **36**(12), 2967–2991 (2003)
10. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: ImageNet: a large-scale hierarchical image database. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition, pp. 248–255. IEEE (2009)
11. Gallardo-Caballero, R., García-Orellana, C., García-Manso, A., González-Velasco, H., Macías-Macías, M.: Independent component analysis to detect clustered microcalcification breast cancers. Sci. World J. **2012**(1), 540457 (2012)
12. Ge, J., et al.: Computer aided detection of clusters of microcalcifications on full field digital mammograms. Med. Phys. **33**(8), 2975–2988 (2006)
13. Giaquinto, A.N., et al.: Breast cancer statistics, 2022. CA: Cancer J. Clin. (2022)
14. Guo, Y., et al.: A new method of detecting micro-calcification clusters in mammograms using contourlet transform and non-linking simplified PCNN. Comput. Methods Programs Biomed. **130**, 31–45 (2016)
15. Hakim, A., Prajitno, P., Soejoko, D.: Microcalcification detection in mammography image using computer-aided detection based on convolutional neural network. In: AIP Conference Proceedings. AIP Publishing (2021)
16. Halling-Brown, M.D., et al.: Optimam mammography image database: a large-scale resource of mammography images and clinical data. Radiol.: Artif. Intell. **3**(1), e200103 (2020). https://medphys.royalsurrey.nhs.uk/omidb/about-omi-db/
17. Hamidinekoo, A., Denton, E., Rampun, A., Honnor, K., Zwiggelaar, R.: Deep learning in mammography and breast histology, an overview and future trends. Med. Image Anal. **47**, 45–67 (2018)
18. 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)

19. Houssein, E.H., Emam, M.M., Ali, A.A., Suganthan, P.N.: Deep and machine learning techniques for medical imaging-based breast cancer: a comprehensive review. Expert Syst. Appl. **167**, 114161 (2021)
20. Karale, V.A., et al.: A screening cad tool for the detection of microcalcification clusters in mammograms. J. Digit. Imaging **32**, 728–745 (2019)
21. Lin, T.Y., Goyal, P., Girshick, R., He, K., Dollár, P.: Focal loss for dense object detection. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 2980–2988 (2017)
22. Lin, T.Y., et al.: Microsoft COCO: common objects in context. In: ECCV 2014, Part V, pp. 740–755. Springer (2014)
23. Litjens, G., et al.: A survey on deep learning in medical image analysis. Med. Image Anal. **42**, 60–88 (2017)
24. Liu, D., Wu, B., Li, C., Sun, Z., Zhang, N.: TrEnD: a transformer-based encoder-decoder model with adaptive patch embedding for mass segmentation in mammograms. Med. Phys. **50**(5), 2884–2899 (2023)
25. Liu, Z., et al.: Swin transformer: hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 10012–10022 (2021)
26. Liu, Z., Mao, H., Wu, C.Y., Feichtenhofer, C., Darrell, T., Xie, S.: A convnet for the 2020s. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 11976–11986 (2022)
27. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101 (2017)
28. Moreira, I.C., Amaral, I., Domingues, I., Cardoso, A., Cardoso, M.J., Cardoso, J.S.: Inbreast: toward a full-field digital mammographic database. Acad. Radiol. **19**(2), 236–248 (2012)
29. Oliver, A., et al.: Automatic microcalcification and cluster detection for digital and digitised mammograms. Knowl.-Based Syst. **28**, 68–75 (2012)
30. Oporto-Díaz, S., Hernández-Cisneros, R., Terashima-Marín, H.: Detection of microcalcification clusters in mammograms using a difference of optimized gaussian filters. In: International Conference Image Analysis and Recognition, pp. 998–1005. Springer (2005)
31. Rehman, K.U., Li, J., Pei, Y., Yasin, A., Ali, S., Mahmood, T.: Computer vision-based microcalcification detection in digital mammograms using fully connected depthwise separable convolutional neural network. Sensors **21**(14), 4854 (2021)
32. Samuelson, F.W., Petrick, N.: Comparing image detection algorithms using resampling. In: 2006 3rd IEEE International Symposium on Biomedical Imaging: Nano to Macro, pp. 1312–1315. IEEE (2006)
33. Savelli, B., Bria, A., Molinara, M., Marrocco, C., Tortorella, F.: A multi-context CNN ensemble for small lesion detection. Artif. Intell. Med. **103**, 101749 (2020)
34. Tan, M., Le, Q.: EfficientNet: rethinking model scaling for convolutional neural networks. In: International Conference on Machine Learning, pp. 6105–6114. PMLR (2019)
35. Tot, T., Gere, M., Hofmeyer, S., Bauer, A., Pellas, U.: The clinical value of detecting microcalcifications on a mammogram. In: Seminars in Cancer Biology, vol. 72, pp. 165–174. Elsevier (2021)
36. Vaswani, A., et al.: Attention is all you need. Advances in Neural Information Processing Systems, vol. 30 (2017)
37. Wei, L., Yang, Y., Nishikawa, R.M., Jiang, Y.: A study on several machine-learning methods for classification of malignant and benign clustered microcalcifications. IEEE Trans. Med. Imaging **24**(3), 371–380 (2005)

38. Wightman, R., Touvron, H., Jégou, H.: ResNet strikes back: an improved training procedure in timm. arXiv preprint arXiv:2110.00476 (2021)
39. Yang, Z., Liu, S., Hu, H., Wang, L., Lin, S.: RepPoints: point set representation for object detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 9657–9666 (2019)
40. Zhu, X., Su, W., Lu, L., Li, B., Wang, X., Dai, J.: Deformable DETR: deformable transformers for end-to-end object detection. arXiv preprint arXiv:2010.04159 (2020)

Unsupervised Feature Matching for Affine Histological Image Registration

Vladislav A. Pyatov and Dmitry V. Sorokin(✉)

Laboratory of Mathematical Methods of Image Processing, Faculty of Computational Mathematics and Cybernetics, Lomonosov Moscow State University, Moscow, Russia
dsorokin@cs.msu.ru

Abstract. One of the most common tasks in histopathology is the visual comparison of the images of successive multiply stained tissue sections. Automatic image registration is crucial to perform this analysis. Although the tissue sections in general undergo non-rigid deformations, the initial linear image alignment impacts the overall registration drastically. However, most of the recent works do not study the linear transformation compensation separately and focus on the non-linear part. In this work, we propose a novel unsupervised feature matching approach for affine registration of histological images. We perform the evaluation on the Automatic Non-rigid Histological Image Registration (ANHIR) dataset and show the supremacy of our method over the existing affine registration approaches in therms of accuracy and robustness. The code is available at https://github.com/VladPyatov/UnFeMa.

Keywords: Histological Imaging · Image Registration · Unsupervised Learning · Vision Transformers

1 Introduction

Image registration aims at establishing spatial correspondences between a pair or a set of images. Being one of the key tasks in biomedical image analysis this research area attracted a lot of attention from the community over the last several decades [6,27]. Even though there are some general methodologies [2,9] that do not impose any specific limitations on the data, each application domain of biomedical image analysis has its own peculiarities that have to be taken into account when solving the registration task.

In digital pathology visual comparison of the successive multiply stained tissue sections is crucial. This requires aligning all images into a common frame, which is also necessary for applications like 3D reconstruction [15] and image fusion [14]. Using an aligned images, pathologists can evaluate expression of multiple markers in a single area. However, tissue processing and image acquisition

The work was financially supported by the Russian Science Foundation under the research project No. 22-41-02002.

A. Antonacopoulos et al. (Eds.): ICPR 2024, LNCS 15313, pp. 34–48, 2025.
https://doi.org/10.1007/978-3-031-78201-5_3

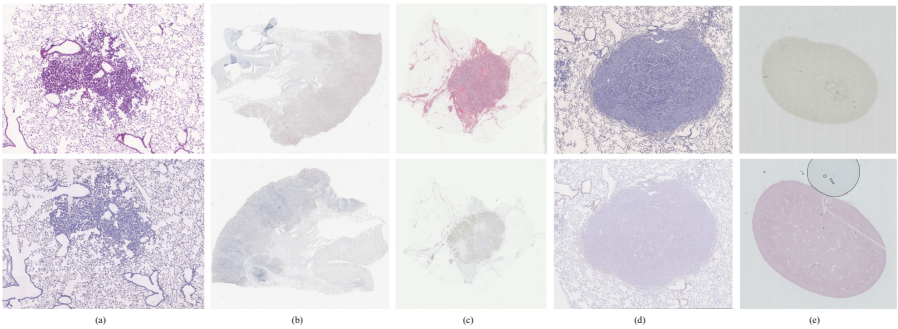


Fig. 1. Examples of the typical histological image pairs from the ANHIR dataset demonstrating the diversity of the data.

procedures cause significant linear transformations and non-linear deformations in the sections. Therefore, image registration is highly demanded.

Histological image registration poses several challenges, as illustrated in Fig. 1:

1. The tissue sections can be globally misaligned: rotations up to 180° and translations (see Fig. 1b, e).
2. The images exhibit variance in appearance due to multiple staining within one sample (see Fig. 1a, c, d, e).
3. The images of different organs can have completely different shape and structure (see Fig. 1a, b, c, d, e).
4. There are repetitive patterns and low textures (see Fig. 1c, e).
5. Section missing and occlusions may occur during tissue section preparation (see Fig. 1b, e).

Thus, the image registration approaches designed for other common medical image modalities like MRI [6], CT [17], or fluorescence microscopy [23] result in imperfect registration or even fail completely on histological data.

In this work, we propose a novel end-to-end unsupervised feature-based affine histological image registration approach and quantitatively compare it to the existing registration methods. To the best of our knowledge, this is the first approach that take advantage of unsupervised learning-based feature matching and explicit estimation of affine transformation parameters. The proposed approach was compared with two classic methods, *i.e.* SIFT [13] and AGH [30], our previous method [16] that uses off-the-shelf LoFTR matches [24], and two deep learning-based approaches DeepHistReg [29] and LARHI [28]. To summarize, the main contributions are as follows:

1. We propose unsupervised feature matching approach for the estimation of affine transformation between a pair of images using the confidence weighting optimization.

2. Based on the proposed approach we developed an end-to-end trainable method for affine histological image registration.
3. Additionally, we present a robust initial alignment algorithm, called Perceptual Search, to find the initial rotation and translation.
4. Our results outperform the existing affine histological image registration methods by a large margin.

2 Related Work

2.1 Learning-Based Affine Image Registration

Affine medical image registration approaches can be divided into two major groups: classical and learning-based. While the former involve an iterative optimization for each image pair [9] or estimate the transformation based on detected features [13], the latter formulate affine registration as a learning problem adopting neural networks of different architectures [28,29,32]. Such formulation allows learning-based methods to hide tuning of hyper-parameters and iterative optimization under the hood of the training procedure. As a consequence learning-based methods perform fast estimation of affine transformation parameters at the inference time usually at the price of arguable registration accuracy.

In DeepHistReg framework [29], the authors propose a simple ResNet-like convolutional neural network that takes as input concatenated source and target images and after series of residual blocks aggregates information within feature maps via global average pooling. This high-dimensional representation is then used to regress affine transformation matrix with one linear layer. The limited receptive field of convolution enables to account for a local misalignment and can be beneficial for deformable registration. However, affine transformation is generally global and brings the maximum benefit compensating for a large displacements.

In contrast to traditional CNNs, Wodzinski and Müller [28] proposed a network architecture that is capable of aggregating global information. The source and target images are first unfolded to a grid of non-overlapping patches and fed into the Siamese feature extraction network. Then the features are concatenated and the global correspondence is extracted with a 3D convolution followed by the MLP for transformation regression. This work pioneered affine registration methods focused on aggregation of the global context, but was fully based on convolutional architecture inheriting its inductive bias.

These methods demonstrate the state-of-the-art performance, but learn to estimate the transformation parameters directly. Hence, the ability to apply these methods to various image resolutions, as well as their robustness, are uncertain. In contrast, we propose to follow classic approaches and decompose learning-based image registration into two parts - feature matching and explicit estimation of transformation parameters.

2.2 Feature Matching

All above mentioned image registration methods are intensity-based. That is, to estimate a transformation they operate on the raw image intensity information. Another approach for affine registration is based on feature matching, the task of establishing correspondences between two images of the same object. Conventional matching approaches consist of three stages - feature detection, feature description, and feature matching. Classic hand-crafted features, such as ORB [19] and SIFT [13], have been widely adopted in keypoint detection and description stages for a long time. Later several works [4,31] proposed learning-based approaches to regress interest point locations and descriptors in a single forward pass of the neural network. Although the aforementioned methods take advantage of learnable features, they use the nearest neighbor search (NNS) to find reliable matches from all interest points detected in the first image to all the interest points in the second image. In contrast, SuperGlue [20] proposed a learning-based approach for local feature matching. Inspired by this work, LoFTR [24] proposed a detector-free design to avoid the drawbacks of feature detectors and directly produce dense feature matches.

Histological images are usually stained with different biomarkers that affect tissue appearance color-wise and texture-wise. Therefore, feature-based formulation of registration problem is more suitable for histological data. Recently, Awan *et al.* [1] proposed affine registration approach based on matching of deep features. Instead of applying feature detection stage, an image is divided into a grid and the feature descriptor is computed for every grid cell. Feature descriptor is formed by deep features extracted from three different layers of a pre-trained VGG-16 network [22]. Inspired by the Transformer's success in modelling long-range dependencies, another recent work [16] leverage LoFTR, pretrained on MegaDepth dataset [10], to perform feature extraction and matching. General limitation of these methods is the exploitation of pretrained models. In contrast, we propose a feature-based affine image registration approach that is end-to-end trainable and thus enables to learn patterns that are specific to histological images.

3 Method

Let $I_S(\boldsymbol{x}) : \mathbb{R}^2 \mapsto \mathbb{R}$ and $I_T(\boldsymbol{x}) : \mathbb{R}^2 \mapsto \mathbb{R}$ be the source and target images, respectively.

Affine image registration aims to estimate an affine transformation $\Lambda : \mathbb{R}^2 \mapsto \mathbb{R}^2$, such that:

$$I_S \circ \Lambda \approx I_T \tag{1}$$

where $I_S \circ \Lambda$ represents I_S transformed by Λ.

We propose a novel feature-based end-to-end learnable affine histological image registration method, that consists of the following parts:

- **Initial alignment with Perceptual Search.** Given image pair (I_S, I_T), Perceptual Search estimates the initial rotation angle $\hat{\varphi}$ to compensate for a large initial misalignment, *e.g.*$\sim 180°$.

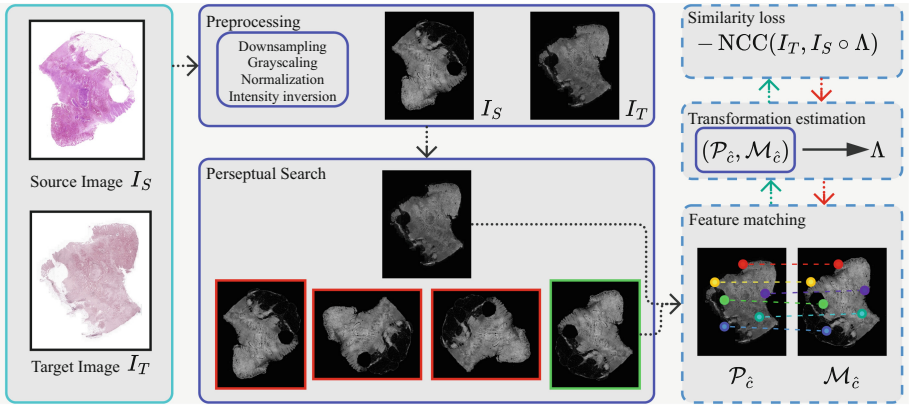


Fig. 2. The overview of the proposed approach.

- **Feature matching with Soft Point refinement.** To establish the coarse-level matches, LoFTR processes the prealigned images and outputs a set of matched points. The obtained matches are then refined with the Soft Points algorithm.
- **Weighted estimation of affine transformation.** The set of matched points serves as an input to the weighted Direct Linear Transformation (DLT) algorithm, that estimates the affine transformation parameters.

Figure 2 shows the overview of the proposed approach. We first introduce the initial alignment with Perceptual Search algorithm in Sect. 3.2. Then, we describe feature matching and Soft Points refinement in Sect. 3.3, and estimation of the affine transformation in Sect. 3.4.

3.1 Preprocessing

The preprocessing algorithm is standard for histological image registration and consists of the following steps.

Since fine-level details are not necessary to find an affine transformation, the first step is to perform downscaling of the originally large images. We choose to downscale images to 512 pixels in the biggest dimension for training for computational efficiency and to 1024 pixels during inference for higher registration accuracy.

The slides of the histological tissues are stained with different dyes during preparation stage. As a result, both global and local color appearance change from slide to slide. This information is essential for biomedical analysis but redundant for affine transformation estimation and in some cases leads to lower registration accuracy [1]. For this reason, we convert the images to grayscale and normalize their intensities. We additionally invert image intensities for convenience, as histological tissues usually appear on the white background.

3.2 Perceptual Search

Most of the existing methods [1,28–30] perform Exhaustive Search - initial alignment step prior to affine registration. This procedure usually involves rough estimataion of translation and rotation parameters. First, the translation vector $\hat{t}$ is estimated based on the centroids of the source and target images. Then, the rotation angle $\hat{\varphi}$ is obtained via the exhaustive search with a predefined step, *e.g.*1°, and normalized correlation coefficient (NCC) similarity measure.

When implemented on GPU, the aforementioned approach is fast, but produces a lot of outliers, *i.e.*pairs that can not be accurately registered on subsequent stages. However, for unsupervised image registration approaches, initial alignment is an important step to avoid overfitting on local shape structure (i.e. when images are misaligned by 180°). Therefore, we propose more precise initial alignment method - Perceptual Search. This procedure is required only to prepare the data for training and thus can be substantially simplified during inference (see Sect. 3.6). Nevertheless, it can be utilized as a standalone initial alignment step.

We do not estimate the translation vector and limit the angle search set to four angles $\Phi = \{0^\circ, 90^\circ, 180^\circ, 270^\circ\}$, since smaller global deformations can be learned well during training. To find the best initial angle $\hat{\varphi} \in \Phi$ we calculate the alignment loss $\mathcal{L}_{\text{al}}$ based on the perceptual loss in the feature space of the pretrained VGG-19 [8] network, replacing L_2 distance with negative NCC

$$\mathcal{L}_{\text{al}}(I_S, I_T) = -\sum_{i,j,c} \text{NCC}\left(V_i^c\left(I_{S_j}\right), V_i^c\left(I_{T_j}\right)\right) \tag{2}$$

where V_i^c is the c^{th} channel of the i^{th} layer of the VGG-19 pretrained network and j represents the resolution. Similar to [21], we use the pyramid of 6 resolutions - 1024×1024, 512×512, 256×256, 128×128, 64×64 and 32×32.

In some cases, the displacement between images can be large after rotation (see Fig. 1b), so we additionally calculate $\mathcal{L}_{\text{al}}(I_S \circ A_\theta, I_T)$, where A_θ denotes affine transformation estimated with matches predicted by LoFTR [24] pretrained on MegaDepth [10] dataset. Therefore, the search loss $\mathcal{L}_{\text{s}}$ and the best initial angle $\hat{\varphi}$ are defined as:

$$\mathcal{L}_{\text{s}}(I_S, I_T) = \min\left(\mathcal{L}_{\text{al}}(I_S, I_T), \mathcal{L}_{\text{al}}(I_S \circ A_\theta, I_T)\right) \tag{3}$$

$$\hat{\varphi} = \underset{\varphi \in \Phi}{\text{argmin}}\, \mathcal{L}_{\text{s}}(I_S \circ R_\varphi, I_T) \tag{4}$$

where R_φ is the rotation transformation through an angle φ about the image center.

3.3 Feature Matching

Preliminaries: LoFTR Given the pair of images ($I_S \in \mathbb{R}^{H_1 \times W_1}, I_T \in \mathbb{R}^{H_2 \times W_2}$), LoFTR utilize feature pyramid network [11] to extract multi-level

features $(\tilde{F}_S, \tilde{F}_T)$ at 1/8 of the original resolution, and $(\hat{F}_S, \hat{F}_T)$ at 1/2 of the original resolution.

Extracted coarse-level features $(\tilde{F}_S, \tilde{F}_T)$ are then passed through the LoFTR module - the interleaving self and cross attention layers - to obtain position and context dependent local features $(\tilde{F}_S^{tr}, \tilde{F}_T^{tr})$.

To find matches, the score matrix $\mathcal{S} \in \mathbb{R}^{\frac{H_S}{8} \cdot \frac{W_S}{8} \times \frac{H_T}{8} \cdot \frac{W_T}{8}}$ between the transformed features is first calculated as:

$$\mathcal{S}(i,j) = \frac{1}{\tau} \cdot \left\langle \tilde{F}_S^{tr}(i), \tilde{F}_T^{tr}(j) \right\rangle \tag{5}$$

where i and j represent coarse-level coordinates of I_S and I_T, respectively. Depending on the context, the coordinates either linearized or cartesian.

A dual-softmax operator [18,26] is then applied to obtain the confidence matrix $\mathcal{P}_c$ of soft mutual nearest neighbor matches:

$$\mathcal{P}_c(i,j) = \text{softmax}(\mathcal{S}(i,\cdot))_j \cdot \text{softmax}(\mathcal{S}(\cdot,j))_i \tag{6}$$

After that, the matches with confidence higher than a threshold of θ_c are refined to the original image resolution with the coarse-to-fine LoFTR module. The matches for refinement $\mathcal{M}_c$ are selected based on the mutual nearest neighbor (MNN) criteria:

$$\mathcal{M}_c = \left\{ (\tilde{i},\tilde{j}) \mid \forall (\tilde{i},\tilde{j}) \in \text{MNN}\left(\mathcal{P}_c\right), \mathcal{P}_c(\tilde{i},\tilde{j}) \geq \theta_c \right\} \tag{7}$$

A dual-softmax matching operator is differentiable and matching confidences $\mathcal{P}_c$ can be directly supervised with the negative log-likelihood loss. However, the MNN criteria is non-differentiable:

$$(\tilde{i},\tilde{j}) \in \text{MNN}\left(\mathcal{P}_c\right) \Leftrightarrow \begin{cases} \tilde{i} = \underset{i}{\text{argmax}}\, \mathcal{P}_c(i,\tilde{j}) \\ \tilde{j} = \underset{j}{\text{argmax}}\, \mathcal{P}_c(\tilde{i},j) \end{cases} \tag{8}$$

Therefore, we can not backpropagate gradients through the coordinates of the matches and perform unsupervised learning of the image registration pipeline. To tackle this problem, we introduce two improvements that make the method end-to-end trainable. First one aims to make the coordinates differentiable. The second is carried out at the stage of transformation estimation and allows to backpropagate directly through the matching confidences $\mathcal{P}_c$.

Soft Points. Given a set of matched pairs, we form a set of soft matches $\mathcal{M}_{\hat{c}}$ by estimating the expected coordinates (Soft Points) for each pair $(\tilde{i},\tilde{j})$:

$$\hat{i} = \mathbb{E}_{p(i \mid \tilde{i},\tilde{j})}\, i; \quad \hat{j} = \mathbb{E}_{p(j \mid \tilde{i},\tilde{j})}\, j \tag{9}$$

where $p(i \mid \tilde{i},\tilde{j})$ is the softmax distribution of scores $\mathcal{S}$ in the local vicinity $U_\varepsilon(\tilde{i})$ of the coordinates $\tilde{i}$ for the pair $(\tilde{i},\tilde{j})$:

$$p(i \mid \tilde{i},\tilde{j}) = \underset{i \in U_\varepsilon(\tilde{i})}{\text{softmax}}(\mathcal{S}(i,\tilde{j})) \tag{10}$$

We also define the soft confidence $\mathcal{P}_c(\hat{i}, \hat{j})$ of the soft match $(\hat{i}, \hat{j})$ as:

$$\mathcal{P}_c(\hat{i}, \hat{j}) = \frac{\left(\mathbb{E}_{p(i \mid \tilde{i}, \tilde{j})} \mathcal{P}_c(i, \hat{j}) + \mathbb{E}_{p(j \mid \tilde{i}, \tilde{j})} \mathcal{P}_c(\hat{i}, j)\right)}{2} \tag{11}$$

The size of the vicinity ε is a hyper-parameter and when it equals 0, the Eq. (9) degenerate into the Eq. (8). We choose to set $\varepsilon = 1$, because this case adds flexibility of refinement of the matched coarse-level coordinates to the original image resolution. Hence, Soft Points serves as a replace for the fine module of LoFTR, as the use of separate network is beneficial with good supervision, but introduces more degrees of freedom and complicates unsupervised learning.

3.4 Transformation Estimation

To estimate the affine transformation Λ, we use Direct Linear Transformation (DLT) algorithm [7], which is reduced to the solution of the following *least squares* minimization problem:

$$x^\star = \underset{\|x\|_2=1}{\operatorname{argmin}} \|Ax\|_2^2, \tag{12}$$

where $x = \operatorname{vec}(\Lambda)$ is the vectorized Λ and $A = A(\mathcal{M}_{\hat{c}})$ is constructed from the matched coordinates.

The approximate solution $\boldsymbol{x}^\star$ of Eq. (12) is the right singular vector corresponding to the smallest singular value of the matrix A [7]. The SVD decomposition of the matrix A can be computed in the differentiable way. Thus the algorithm enables to backpropagate gradients through the solution $\boldsymbol{x}$ and learn the matches $\mathcal{M}_{\hat{c}}$ if the matrix A depends on the parameters of the matching network. In our case A depends on the Soft Points. However, because of the locality ($\varepsilon = 1$), the Soft Points propagate gradients only through the local vicinity U_ε. The increasing of ε, however, leads to unstable optimization. With this observations, we propose to additionally backpropagate gradients directly through confidences $\mathcal{P}_c(\hat{i}, \hat{j})$, by solving a confidence weighted modification of this system:

$$x^\star = \underset{\|x\|_2=1}{\operatorname{argmin}} \|WAx\|_2^2, \tag{13}$$

where $W \in \mathbb{R}^{2N \times 2N}$ is the diagonal matrix, which weighs each pair of equations corresponding to the match pair $(\hat{i}, \hat{j})$ with the confidence $\mathcal{P}_c(\hat{i}, \hat{j})$.

The intuition of the weighting is that more confident matches should have a higher influence on the system and its solution. More importantly, this technique allows to backpropagate gradients directly through confidences $\mathcal{P}_c$ and learn the matches $\mathcal{M}_{\hat{c}}$ alleviating the problem of the non-differentiable MNN criteria.

The algorithm of weighted estimation of the affine transformation is implemented as a separate module that is fully differentiable and makes image registration pipeline end-to-end trainable.

3.5 Training Loss

After the estimation of the affine transformation matrix Λ, we use the negative local NCC to calculate the training objective $\mathcal{L}$ between the target image I_T and the transformed source image $I_S \circ \Lambda$:

$$\mathcal{L} = -\frac{1}{|P|} \sum_{\boldsymbol{p} \in P} \mathrm{NCC}\left(I_T(\boldsymbol{p}), I_S \circ \Lambda(\boldsymbol{p})\right) \tag{14}$$

where P is a set of non-overlapping patches of the size 32×32. Thus, learning of $\mathcal{P}_c(\hat{i}, \hat{j})$ is organized in unsupervised fashion and the matches are optimized indirectly through maximization of the similarity between I_T and $I_S \circ \Lambda$ images after registration.

3.6 Implementation Details

In our approach we utilize LoFTR with QuadTree attention [25] for the feature matching stage. We initialize its weights from the model pretrained on the MegaDepth [10] dataset and train it on the histological data in the proposed image registration pipeline. The model is trained 32 epochs on 2 NVIDIA RTX A6000 with batch size 16. We adopt the AdamW optimizer [12] and start with learning rate $1 \cdot 10^4$ reducing it by half every 4 epochs.

At the inference stage we follow [16] to make registration resistant to the global misalignment and check four possible angles $\varphi \in \{0°, 90°, 180°, 270°\}$. Namely, for each φ we predict the matches between $I_S \circ R_\varphi$ and I_T and then estimate the affine transformation Λ with RANSAC [5] as a robust estimator to discard outliers. The resulting transformation is the composition of rotation R_φ and corresponding affine transformation Λ that gives the highest NCC similarity.

4 Experiments

4.1 Dataset

We benchmark the proposed approach on the Automatic Non-rigid Histological Image Registration (ANHIR) dataset [3], which includes the histological images of eight different tissue types with approximate size of $10k \times 10k$ pixels. The images are organized into 49 sets of spatially close tissue sections that can be meaningfully registered. Each image set contains 3 to 9 images. In total, the dataset contains 355 images with 18 different stains, forming 481 image pairs. The images were manually marked with the pairs of landmarks to enable evaluation of the registration performance. The dataset is divided in two subsets. The Training subset consists of the 230 image pairs with corresponding landmarks for the source and target images. The Evaluation subset of 251 image contains landmarks only for the source images and serves for the server-side evaluation.

It is important to note that landmarks were not used during model training, and the proposed approach is fully unsupervised. The division of available data

Table 1. Median Median (MMrTRE), Average Median (AMrTRE), Median Average (MArTRE), and Average Average (AArTRE) errors for the evaluated methods on the Train and Evaluation datasets. (t) and (e) denote that method was trained on the Training and Evaluation datasets, respectively. Best results highlighted with bold, second best results are underlined.

Method	Training				Evaluation			
	MMrTRE	AMrTRE	MArTRE	AArTRE	MMrTRE	AMrTRE	MArTRE	AArTRE
Initial	0.039315	0.059856	0.040256	0.060415	0.051002	0.121860	0.052610	0.120204
SIFT [13]	0.013733	0.095091	0.015747	0.103877	0.023117	0.166627	0.026639	0.163326
AGH [30]	0.003792	0.009400	0.005349	0.010613	0.003928	**0.008392**	0.005070	**0.009624**
TAHIR [16]	0.003639	0.009256	0.004800	0.010691	0.003191	0.010595	0.004335	0.012196
DeepHistReg [29] (t)	–	–	–	–	0.016862	0.040922	0.019101	0.042921
DeepHistReg [29] (e)	0.017437	0.051237	0.018770	0.051990	–	–	–	–
LARHI [28] (t)	–	–	–	–	0.018492	0.042270	0.020563	0.044280
LARHI [28] (e)	0.019787	0.053931	0.020094	0.054872	–	–	–	–
Proposed (t)	–	–	–	–	**0.003027**	0.010573	**0.004162**	0.012429
Proposed (e)	**0.003433**	**0.008212**	**0.004629**	**0.009636**	–	–	–	–

into training and evaluation was provided by ANHIR challenge organizers and related only to the ability to run the evaluation locally or on the challenge server-side. Thus, we evaluated the learning-based approaches in two scenarios: first, training on the training data and evaluating on the evaluation data, and second, training on the evaluation data and evaluating on the training data (see Table 1).

4.2 Evaluation Metrics

We use the aggregation metrics based on *relative Target Registration Error* (rTRE) and *Robustness* introduced in [3]. The (rTRE) is the Euclidean distance between the coordinates of the landmarks in the transformed source I_S and target I_T images

$$\mathrm{rTRE}_l^{S,T} = \frac{\|\mathbf{x}_l^S - \mathbf{x}_l^T\|_2}{d_S} \tag{15}$$

normalized by the length of the image diagonal d_S.

The *Robustness* is defined as the relative number of successfully registered landmarks, *i.e.* those for which the registration error decreases, $\mathrm{rTRE}_l < \mathrm{rIRE}_l$. Here, rIRE_l is the *relative Initial Registration Error* which is computed as rTRE for unregistered pair of images. Finally, the *Robustness* is defined as

$$R^{T,S}(m) = \frac{|K^{T,S}|}{|L^T|} \tag{16}$$

where $K^{T,S} = \left\{(\mathbf{x}_l^S, \mathbf{x}_l^T) : \mathrm{rTRE}_l < \mathrm{rIRE}_l\right\} \subseteq L^T$ is the set of successfully registered landmarks, and L^T is the set of target image landmarks.

For quantitative evaluation, we utilize Median Median (MMrTRE), Average Median (AMrTRE), Median Average (MArTRE), and Average Average (AArTRE) rTRE, as well as Median (MR) and Average (AR) robustness. It

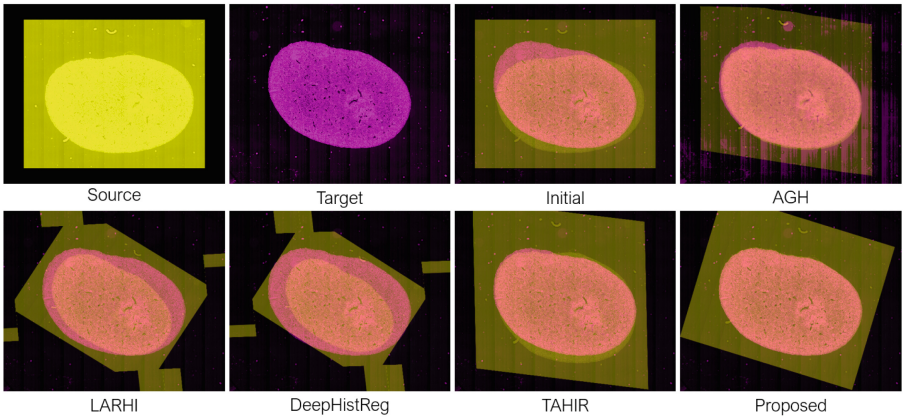


Fig. 3. The visualization of the registration results. Upper row from left to right: initial source image, initial target image, initial images overlay, images registered by AGH; bottom row from left to right: images registered by LARHI, images registered by DeepHistReg, images registered by TAHIR, image registered by the proposed method.

is important to note that for ANHIR challenge rankings, the MMrTRE and MR metrics were considered primary. For further details on evaluation metrics, please refer to [3].

4.3 Results and Comparison

We report the performance of the proposed method in Table 1. For the fair comparison we reproduced the results of several top-performing methods. All the results were produced with the server-side evaluation system provided by the ANHIR challenge organizers.

First, we compare the proposed approach to the classic feature-based affine image registration methods, SIFT [13] with Exhaustive Search for initial alignment, AGH [30] which utilizes several descriptors with gradient-based optimization, and TAHIR [16] that use RANSAC on top of the matches predicted by the MegaDepth pretrained LoFTR matching network. We also compare our method to the learning-based methods that directly predict the transformation matrix. The first is the affine part of the DeepHistReg [29] non-rigid image registration framework, that uses ResNet convolutional neural network. The second approach LARHI [28] takes advantage of the convolutional attention mechanism and patch-wise feature extraction.

The results show, that the proposed method outperforms the other approaches by a large margin. MMrTRE and MArTRE metrics demonstrate high registration accuracy of the proposed approach, while AMrTRE and AArTRE indicates the less number of outliers compared to the other methods.

Table 2. Median (MR) and average (AR) robustness for the evaluated methods on the Training and Evaluation datasets. (t) and (e) denote that a method was trained on the Training and Evaluation parts of the data, respectively

Method	Training		Evaluation	
	MR	AR	MR	AR
Initial	0.705	0.677	0.660	0.659
SIFT [13]	0.917	0.651	0.972	0.730
AGH [30]	**1.0**	**0.958**	**1.0**	0.965
TAHIR [16]	**1.0**	**0.958**	**1.0**	0.957
DeepHistReg [29] (t)	–	–	0.960	0.820
DeepHistReg [29] (e)	0.912	0.754	–	–
LARHI [28] (t)	–	–	0.961	0.815
LARHI [28] (e)	0.864	0.712	–	–
Proposed (t)	–	–	**1.0**	**0.966**
Proposed (e)	**1.0**	0.957	–	–

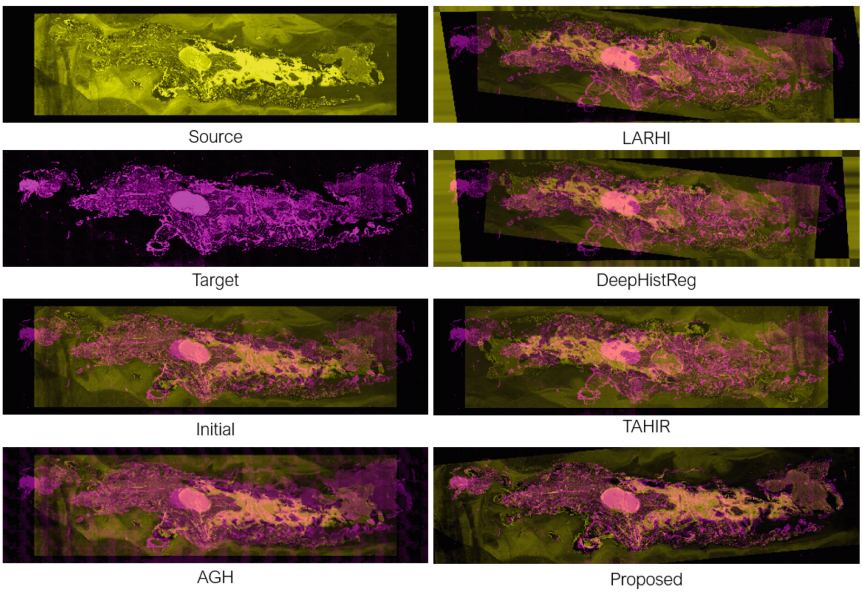


Fig. 4. The visualization of the registration results. Left column from top to bottom: initial source image, initial target image, initial images overlay, images registered by AGH; right column from top to bottom: images registered by LARHI, images registered by DeepHistReg, images registered by TAHIR, image registered by the proposed method.

The reported *Robustness* results (see Table 2) demonstrate that the proposed method not only outperforms the existing approaches, but also behaves more stable than the other learning-based methods.

The visual comparison of the registration results are depicted in Figs. 3 and 4. In Fig. 3, the tissue has nearly elliptic shape which turned out to be hard for affine registration. Namely, AGH and TAHIR methods resulted in strong shear, while LARHI and DeepHistReg scaled down the source image. In Fig. 4, the proposed approach outperforms other methods, evident in the precise alignment of the solid elliptical-shaped tissue fragment at the center of the images.

5 Conclusion

In this paper, we proposed an unsupervised feature matching approach for end-to-end trainable affine histological image registration. The method relies on the improved feature matching with Soft Point refinement and utilizes the weighted Direct Linear Transformation to estimate the affine transformation parameters. Additionally, we introduced the perceptual initial alignment step that enable to compensate for the large rotations and translations prior to the training step. The results on the common publicly available benchmark dataset ANHIR [3] demonstrated that the proposed approach outperforms the existing affine histological image registration methods in both accuracy and robustness.

References

1. Awan, R., Raza, S.E.A., Lotz, J., Weiss, N., Rajpoot, N.: Deep feature based cross-slide registration. Comput. Med. Imaging Graph. **104**, 102162 (2023)
2. Balakrishnan, G., Zhao, A., Sabuncu, M.R., Guttag, J., Dalca, A.V.: VoxelMorph: a learning framework for deformable medical image registration. IEEE Trans. Med. Imaging **38**(8), 1788–1800 (2019)
3. Borovec, J., et al.: ANHIR: automatic non-rigid histological image registration challenge. IEEE Trans. Med. Imaging **39**(10), 3042–3052 (2020)
4. DeTone, D., Malisiewicz, T., Rabinovich, A.: SuperPoint: self-supervised interest point detection and description. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, pp. 224–236 (2018)
5. Fischler, M.A., Bolles, R.C.: Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. Commun. ACM **24**(6), 381–395 (1981). https://doi.org/10.1145/358669.358692
6. Fu, Y., Lei, Y., Wang, T., Curran, W.J., Liu, T., Yang, X.: Deep learning in medical image registration: a review. Phys. Med. Biol. **65**(20), 20TR01 (2020). https://doi.org/10.1088/1361-6560/ab843e
7. Hartley, R.I., Zisserman, A.: Multiple View Geometry in Computer Vision, 2nd edn. Cambridge University Press (2004). ISBN 0521540518
8. Johnson, J., Alahi, A., Fei-Fei, L.: Perceptual losses for real-time style transfer and super-resolution. In: ECCV 2016, Part II, pp. 694–711. Springer, Cham (2016)
9. Klein, S., Staring, M., Murphy, K., Viergever, M.A., Pluim, J.P.: Elastix: a toolbox for intensity-based medical image registration. IEEE Trans. Med. Imaging **29**(1), 196–205 (2009)

10. Li, Z., Snavely, N.: MegaDepth: learning single-view depth prediction from internet photos. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2041–2050 (2018)
11. Lin, T.Y., Dollar, P., Girshick, R., He, K., Hariharan, B., Belongie, S.: Feature pyramid networks for object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2017)
12. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. In: International Conference on Learning Representations (2019)
13. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. Int. J. Comput. Vision **60**(2), 91–110 (2004)
14. Obando, D.F.G., Frafjord, A., Øynebråten, I., Corthay, A., Olivo-Marin, J.C., Meas-Yedid, V.: Multi-staining registration of large histology images. In: 2017 IEEE 14th International Symposium on Biomedical Imaging (ISBI 2017), pp. 345–348 (2017)
15. Pichat, J., Iglesias, J.E., Yousry, T., Ourselin, S., Modat, M.: A survey of methods for 3D histology reconstruction. Med. Image Anal. **46**, 73–105 (2018)
16. Pyatov, V.A., Sorokin, D.V.: TAHIR: transformer-based affine histological image registration. In: Rousseau, J.J., Kapralos, B. (eds.) ICPR 2022, pp. 541–552. Springer, Cham (2023)
17. Rigaud, B., et al.: Deformable image registration for radiation therapy: principle, methods, applications and evaluation. Acta Oncol. **58**(9), 1225–1237 (2019)
18. Rocco, I., Cimpoi, M., Arandjelović, R., Torii, A., Pajdla, T., Sivic, J.: Neighbourhood consensus networks. In: Advances in Neural Information Processing Systems, vol. 31 (2018)
19. Rublee, E., Rabaud, V., Konolige, K., Bradski, G.: ORB: an efficient alternative to SIFT or SURF. In: 2011 International Conference on Computer Vision, pp. 2564–2571 (2011). https://doi.org/10.1109/ICCV.2011.6126544
20. Sarlin, P.E., DeTone, D., Malisiewicz, T., Rabinovich, A.: SuperGLUE: learning feature matching with graph neural networks. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2020)
21. Siarohin, A., Woodford, O.J., Ren, J., Chai, M., Tulyakov, S.: Motion representations for articulated animation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 13653–13662 (2021)
22. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. In: International Conference on Learning Representations (2015)
23. Sorokin, D.V., Tektonidis, M., Rohr, K., Matula, P.: Non-rigid contour-based temporal registration of 2D cell nuclei images using the navier equation. In: 2014 IEEE 11th International Symposium on Biomedical Imaging (ISBI), pp. 746–749 (2014). https://doi.org/10.1109/ISBI.2014.6867978
24. Sun, J., Shen, Z., Wang, Y., Bao, H., Zhou, X.: LoFTR: detector-free local feature matching with transformers. In: Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition (CVPR), pp. 8922–8931 (2021)
25. Tang, S., Zhang, J., Zhu, S., Tan, P.: Quadtree attention for vision transformers. In: International Conference on Learning Representations (2022)
26. Tyszkiewicz, M., Fua, P., Trulls, E.: DISK: learning local features with policy gradient. In: Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., Lin, H. (eds.) Advances in Neural Information Processing Systems, vol. 33, pp. 14254–14265. Curran Associates, Inc. (2020)

27. Viergever, M.A., Maintz, J.A., Klein, S., Murphy, K., Staring, M., Pluim, J.P.: A survey of medical image registration - under review. Med. Image Anal. **33**, 140–144 (2016). https://doi.org/10.1016/j.media.2016.06.030. 20th anniversary of the Medical Image Analysis journal (MedIA)
28. Wodzinski, M., Müller, H.: Learning-based affine registration of histological images. In: International Workshop on Biomedical Image Registration, pp. 12–22 (2020)
29. Wodzinski, M., Müller, H.: Deephistreg: Unsupervised deep learning registration framework for differently stained histology samples. Comput. Methods Programs Biomed. **198**, 105799 (2021)
30. Wodzinski, M., Skalski, A.: Multistep, automatic and nonrigid image registration method for histology samples acquired using multiple stains. Phys. Med. Biol. **66** (2020). https://doi.org/10.1088/1361-6560/abcad7
31. Yi, K.M., Trulls, E., Lepetit, V., Fua, P.: LIFT: learned invariant feature transform. In: ECCV 2016, Part VI, pp. 467–483. Springer, Cham (2016)
32. Zhao, S., Lau, T., Luo, J., Chang, E.I.C., Xu, Y.: Unsupervised 3D end-to-end medical image registration with volume tweening network. IEEE J. Biomed. Health Inform. **24**(5), 1394–1404 (2020). https://doi.org/10.1109/JBHI.2019.2951024

Towards Out-of-Distribution Detection for Breast Cancer Classification in Point-of-Care Ultrasound Imaging

Jennie Karlsson[1(✉)], Marisa Wodrich[1], Niels Christian Overgaard[1], Freja Sahlin[1], Kristina Lång[2,3], Anders Heyden[1], and Ida Arvidsson[1]

[1] Centre for Mathematical Sciences, Lund University, Lund, Sweden
jennie.karlsson@math.lth.se
[2] Department of Translational Medicine, Division of Diagnostic Radiology, Lund University, Lund, Sweden
[3] Unilabs Mammography Unit, Skåne University Hospital, Malmö, Sweden

Abstract. The use of deep learning for classification tasks has shown great potential in medical applications. In critical domains as such, it is of high interest to have trustworthy algorithms which are able to tell when a reliable assessment cannot be guaranteed. Hence, detecting out-of-distribution (OOD) samples is a crucial step towards building a safe classifier. Following a previous study, showing that it is possible to classify breast cancer in point-of-care ultrasound (POCUS) images, this study investigates out-of-distribution (OOD) detection. Three different OOD detection methods were implemented and evaluated in this study: softmax score, multi-level energy score and deep ensembles. As in-distribution training data both standard ultrasound images and POCUS images were used and a separate POCUS data set was used for testing. All OOD detection methods were evaluated on three different OOD data sets, which are a mixture of synthetic data and real ultrasound data that represent different use cases for which OOD detection in automatic breast cancer classification is needed, covering a range of simple OOD cases, ultrasound images of poor quality and ultrasound images of non-breast tissue. The results show that the softmax score is inferior compared to the other methods at detecting OOD samples. The multi-level energy score performs superior on two of the OOD data sets. The deep ensembles perform superior on the OOD data set containing ultrasound images of poor quality with a 95% confidence interval for the area under the receiver operating characteristic curve of 97.2%–98.5%.

Keywords: Out-of-distribution detection · breast cancer classification · point-of-care ultrasound

1 Introduction

Deep learning has achieved promising results assessing different types of medical images [1]. However, for safe deployment and trustworthiness, the algorithms

A. Antonacopoulos et al. (Eds.): ICPR 2024, LNCS 15313, pp. 49–63, 2025.
https://doi.org/10.1007/978-3-031-78201-5_4

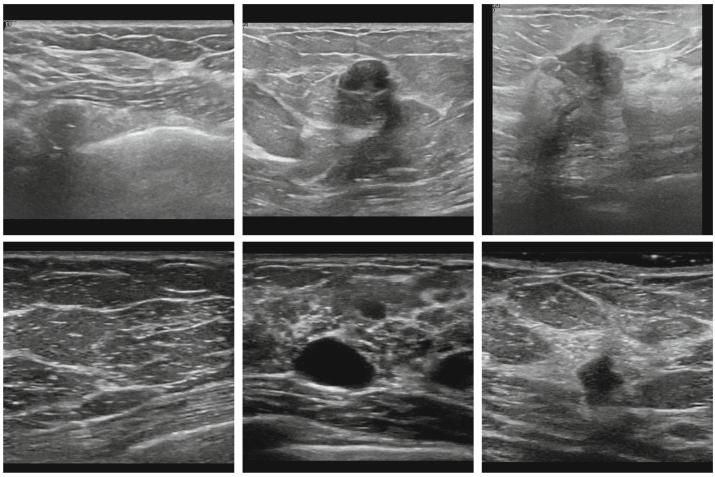


Fig. 1. US (top) and POCUS (bottom) images capturing normal tissue, benign and malignant lesions (from the left to right).

should be able to tell when they cannot make reliable assessments. This includes detecting out-of-distribution (OOD) data, which comprise data samples that the algorithm has not learned how to interpret. Progress has been made in this area in recent years with numerous methods being suggested for OOD detection [2], i.e. separating OOD samples from in-distribution (ID) samples. In this study we explore different methods for OOD detection, including uncertainty-based ones [3], with the aim of being used in a tool for breast cancer classification in point-of-care ultrasound (POCUS) images.

Breast cancer is the most common type of cancer amongst women worldwide [4]. Detecting breast cancer at an early stage improves patient outcome both in terms of mortality and morbidity, but access to diagnostics is lacking in many low- and middle-income countries [5,6]. Poor access to diagnostics is due to limited resources both in terms of medical equipment and trained personnel. A possible method to provide a timely diagnosis of breast cancer in low-resource settings is using POCUS by minimally trained examiners with a deep-learning based algorithm as decision support. Examples of standard ultrasound (US) and POCUS images capturing normal breast tissue, benign and malignant breast lesions are shown in Fig. 1.

In a previous study a convolutional neural network (CNN) has been used to classify breast cancer in POCUS images with good performance [7]. However, the CNN was not suited to decide whether a prediction should be made or not. An unsuitable image should ideally be detected as an OOD sample and no prediction should be made. Three reasons for an image being unsuitable are (1) poor quality, (2) non-breast tissue and (3) rare lesion. The first case includes images of poor quality which can occur due to numerous factors, resulting in artefacts such as shadows and noise. Such factors can be poor transmission

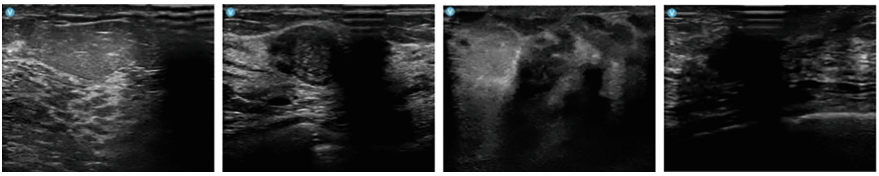


Fig. 2. Examples of poor quality POCUS images that are unsuitable for the classifier to predict. The images contain noise, artefacts and shadows.

capabilities due to a lack of ultrasonic gel during examination, applying too little pressure on the tissue or not holding the probe steady. Examples of POCUS images of poor quality are shown in Fig. 2. The second case comprises images exclusively capturing other structures than breast tissue, for example bones or arteries. Since the CNN is specialized in classifying breast tissue it is neither suitable nor safe to trust its prediction on other structures. The third case are images which are unsuitable due to comprising lesions which the CNN is not familiar with, for example a rare type of cancer.

Here we extend the work done in [7] by exploring and evaluating three different OOD detection methods: softmax score, multi-level energy score and deep ensembles, using both variance- and entropy-based uncertainties. The methods have been evaluated thoroughly on three OOD datasets including both images of poor quality due to noise and images of non-breast tissue. The distributions of OOD scores were analyzed and cancer detection performance at different thresholds for the OOD detection was evaluated. This approach should enable the CNN detecting unsuitable images during inference, which is crucial for safe usage in a real world setting.

2 Theory

2.1 Softmax Score

The probabilities outputted from the softmax activation function after the last layer of a neural network have previously been used for OOD detection [8]. The idea is that samples with low softmax scores for the predicted class are identified as OOD. However, an issue with softmax score is that the scores for the classes need to add up to one, even when none of the classes fits the sample. Thus, softmax score might not be the best suited for OOD detection, but is included in this study as a baseline.

2.2 Energy Score

Energy score is a post-hoc OOD detection method using the logits, i.e. the unscaled outputs from the network before the softmax activation. By looking at the logits, the issue of the softmax score requirement to add up to one is avoided.

The method was proposed in [9] and does not require any retraining of the neural network. For input x and network $f(x)$, the energy score can be expressed as

$$E(x; f) = -T \cdot \log \sum_{i=1}^{K} e^{f_i(x)/T}, \tag{1}$$

where the logit for class label i and input x is denoted by $f_i(x)$, K denotes the total number of class labels and the temperature T is a hyperparameter.

A related work to energy score is multi-level out-of-distribution detection (MOOD) [10]. This method uses adjusted energy score but instead of just analysing the score in the end of the network, it analyses multiple exits that are added throughout the network. The idea is that obvious OOD samples should be detected earlier in the network and more complex samples should be detected further down in the network.

2.3 Deep Ensembles

Deep neural network ensembling [11] is a method to improve the generalizability and reliability of a network by independently training multiple networks on the same problem. The combined output makes it possible to calculate uncertainties by looking at the differences between the separate predictions. These uncertainties can be used for OOD detection. For OOD data samples, the uncertainty should be high as opposed to ID data, where the uncertainty should be small [12]. In the present study, an average ensemble is used, which makes a prediction based on the average prediction from the N ensemble members

$$F_c(x) = \frac{1}{N} \sum_{n=1}^{N} f_c^{(n)}(x), \tag{2}$$

where $F_c(x)$ is the prediction of ensemble F for class c and $f_c^{(n)}$ is the n-th ensemble member's prediction for that class.

Measuring the uncertainty in an ensemble can be achieved in various ways. In this study we include variance-based uncertainty, weighted variance-based uncertainty and entropy-based epistemic uncertainty. The variance-based uncertainty, $\mathcal{U}_{var}$, is calculate as the sum of the variance of the ensemble for each class c out of in total K classes,

$$\mathcal{U}_{var} = \sum_{c=1}^{K} \mathrm{Var}[F_c(x)] = \sum_{c=1}^{K} \frac{1}{N} \sum_{n=1}^{N} (f_c^{(n)}(x) - F_c(x))^2. \tag{3}$$

The weighted variance-based uncertainty, $\mathcal{U}_{weightedVar}$, additionally weights each class variance by the mean, consequentially taking the variance for the predicted class higher into account,

$$\mathcal{U}_{weightedVar} = \sum_{c=1}^{K} F_c(x) \frac{1}{N} \sum_{n=1}^{N} (f_c^{(n)}(x) - F_c(x))^2. \tag{4}$$

Table 1. The sizes of the ID data sets.

	POCUS		US
	Train	Test	Train
Normal	463	284	386
Benign	173	131	254
Malignant	178	116	520
Total	814	531	1160

The entropy-based epistemic uncertainty, $\mathcal{U}_{epi}$, is defined as,

$$\mathcal{U}_{epi} = -\sum_{c=1}^{K} F_c(x)\log F_c(x) + \frac{1}{N}\sum_{n=1}^{N}\sum_{c=1}^{K} f_c^{(n)}(x)\log f_c^{(n)}(x). \tag{5}$$

Epistemic uncertainty is the model uncertainty, which stems from the model's lack of knowledge [13].

3 Data

3.1 In-Distribution Data

The ID data consists of POCUS images capturing breast tissue collected with a GE Vscan air CL probe [14] at Skåne University Hospital, Malmö, Sweden. This data set contains images of normal tissue as well as benign and malignant lesions, see Fig. 1. The data was split into training and test set containing 814 and 531 images respectively.

In addition to the POCUS training set, a standard ultrasound (US) data set containing 1160 images of breast tissue was also used for training. These images were labeled as normal, benign or malignant and were collected with Logiq E9 and Logiq E10 ultrasound machines at Skåne university hospital, Malmö, Sweden. The sizes and division of the ID data sets are shown in Table 1.

3.2 Out-of-Distribution Data

To evaluate the performance in terms of OOD detection, three different OOD test data sets were used: MNIST (test set), CorruptPOCUS and CCA. The MNIST test set consists of 10 000 images of handwritten digits [15] and was used as a baseline to make sure that the OOD detection could handle simple cases of wrong input data. The CorruptPOCUS and CCA data sets were chosen as realistic OOD ultrasound data and cover different use cases. The CorruptPOCUS data set consists of POCUS images of poor quality, which was the first case of unsuitable images mentioned in Sect. 1. The images were generated by distorting each of the 531 images in the POCUS test set by randomly adding dark areas, blur and noise to simulate the issues illustrated by the images in Fig. 2.

The CCA data set contains 84 ultrasound images capturing the common carotid artery [16], therefore comprising images of non-breast structure, the second case of unsuitable images. Covering this case is especially important due to the potential misuse of our system for other diagnostic purposes for which it has not been trained. The third case of unsuitable images was data which the CNN is not familiar with, such as rare types of cancer. To the authors' knowledge, no such data set exists publicly and due to the complexity of collecting such a dataset, this use case was excluded in this paper. Examples of images from the three OOD data sets can be seen in Fig. 3.

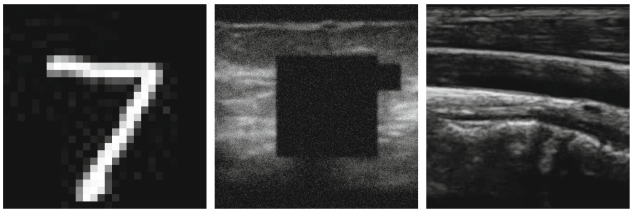


Fig. 3. Example of images from the OOD data sets MNIST, CorruptPOCUS and CCA (left to right).

4 Methods

4.1 Classification Network

The classification network used in this study is based on the architecture presented in [7]. It was implemented as a CNN consisting of five convolutional layers followed by two fully connected layers. The input to the network consisted of one channel images of size 180×180. The kernel sizes of the five convolutional layers were set to 3×3 with 32, 64, 128, 128 and 128 kernels respectively. Each convolutional layer was followed by a ReLU activation, 20% dropout and finally a max pooling layer with pool size 2×2 and stride 2. The final max pooling layer was followed by 50% dropout and the output was flattened and used as input into two fully connected layers of sizes 512 and 3. The ReLU activation and 50% dropout was used between the fully connected layers and the softmax activation was used after the final layer. If nothing else is stated, the following settings of training parameters were used for the classification network and all OOD detection methods. The Adam optimizer was used with a learning rate of 0.0001, the batch size was set to 32 and the network was trained for 50 epochs. The three different classes were weighted to have equal influence during the training of the network. Spatial data augmentation was used randomly for each image and epoch during the training phase to increase the variability within the ID training data. The augmentation was done by applying zoom, shear transform, vertical shift or horizontal shift to the images, all within a range of 10%. Randomly flipping the images horizontally was also included in the augmentation. The architecture of the classification network was used for all experiments with some modifications depending on the method.

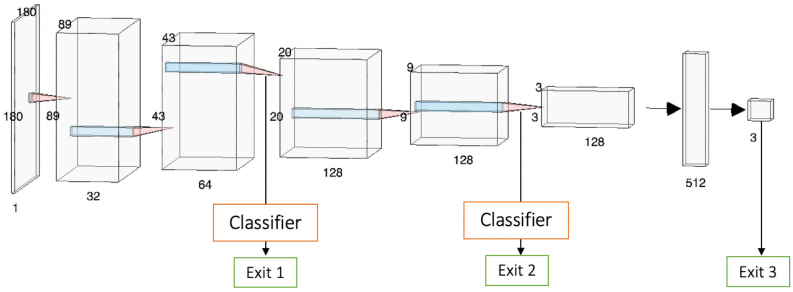


Fig. 4. Architecture of the classification network with additional classifiers and exits. The scheme was created in NN-SVG [17].

4.2 Softmax Score

The softmax score was used as a baseline for the OOD detection methods. The classification network mentioned above was used directly as described in the previous section and the probabilities were obtained from the output of the network. Samples with probabilities lower than a set threshold were detected as OOD.

4.3 Multi-level Energy Score

The energy score was implemented according to Eq. (1), but the classification network was modified to have three exits and the energy score was obtained at multi-levels inspired by the idea behind MOOD. The modified architecture of the classification network is displayed in Fig. 4. The classifiers added to the first two exits consisted of one convolutional layer with 128 kernels of size 3×3, a ReLU activation, a 2×2 max pooling with stride 2 and a fully connected layer of size three without any activation. After trying out different values for the temperature parameter T based on the ID data, it was set to 0.001. The classification network was trained with three categorical cross-entropy losses, using the outputs from the first two exits and the final softmax output. The three losses were weighted 0.5, 0.5, 1, making the last exit influence the training the most. For OOD detection, thresholds were found for each exit and if a sample had higher energy score than all three thresholds it was labeled as ID, otherwise OOD. The thresholds were chosen with the constraint of equal fraction of data detected as OOD for each exit. For cancer classification task only the final exit, i.e., the output of the network was used.

4.4 Deep Ensemble

For the deep ensemble, 20 independent models with the above described classification network architecture (see Sect. 4.1) were trained. In order to diversify the

models 0–15% of the training data was randomly left out for each of model, followed by setting some of the training hyperparameters randomly. Table 2 specifies the training parameters that were randomly set and the corresponding setting options.

Table 2. Possible training parameters for the deep ensemble models.

Parameter	Options
Random training split	0–15%
Learning rate	0.0001–0.001
Optimizer	Adam or RMSprop
Epochs	25–85
Batch size	8, 16, 32, 64 or 128

As the measure for uncertainty for the deep ensemble three different methods were used: variance-based uncertainty, weighted variance-based uncertainty and entropy-based epistemic uncertainty. For the variance-based uncertainty it was used both with and without being weighted before summation, see Eq. 3 and Eq. 4. The entropy-based epistemic uncertainty was used according to Eq. 5.

4.5 Experiments and Metrics

In a first experiment, all methods were evaluated for the purpose of detecting the OOD samples from the three different OOD data set. The aim was to separate POCUS test set (ID data) from OOD data solely based on the score obtained from the OOD methods. In a second experiment, the methods were tested on the POCUS test set for finding OOD samples within that data set. Excluding samples that are detected as OOD should ideally improve the classification performance on the rest of the data, ultimately making the predictions more trustworthy. Three types of metrics were used to evaluate the different methods: receiver operating characteristic (ROC) curve, area under the ROC curve (AUC) and false positive rate (FPR).

For the first experiment, both ROC and AUC were evaluated, using the POCUS test set as ID and the OOD sets as OOD. The FPR was evaluated at the OOD detection thresholds of 95% and 80% true positive rate (TPR) for the POCUS training data, i.e. 95% or 80% of the data was detected as ID, referred to as FPR95 and FPR80 respectively. The multi-level energy score was evaluated both for each exit separately and combined. For statistical testing, the 95% confidence interval (CI) for the AUC (AUC 95% CI) was calculated using bootstrapping on the ID test set 1000 times.

For the second experiment, the classification performance into normal and benign (non-cancerous) versus malignant (cancerous) on the POCUS test set was measured using AUC and AUC 95% CI. The performance was measured

when using the whole test set, as well as when excluding samples based on the OOD detection thresholds described for the previous experiment. Given these thresholds, 5% or 20% of the POCUS training data would be left out, referred to as leave-out-rate. Only the remaining samples were included in the classification performance evaluation. Statistical significant classification improvement was computed using the Mann-Whitney U-test, with p-values less than 0.05 considered significant.

5 Results

The AUC and corresponding CI, along with FPR80 and FPR95 were evaluated for all OOD detection methods for each OOD data set, see Table 3. For the MNIST and CCA data sets, the multi-level energy score has the best performance with an AUC of 98.1% and 78.5% respectively. The CorruptPOCUS data set is best detected using the deep ensemble with variance-based uncertainty, with an AUC of 97.9%. For all the methods and OOD data sets, except for the softmax score on MNIST, the false positive rate decreases when using FPR80 compared to FPR95.

The ROC curves for each method and OOD dataset can be seen in Fig. 5, showing the relation between false positive rate and true positive rate. The softmax score performs the worst for all datasets and false positive rates. The multi-level energy score performs the best for MNIST at all false positive rates and CCA for low false positive rates. All the deep ensemble methods perform similar with best performance for CorruptPOCUS at all false positive rates.

The individual results for OOD detection using energy scores from the different exits can be seen in Table 4. The performance of the separate energy scores from the three exits varies for the different OOD data sets, with MNIST being best detected in exit 2 and 3, and CorruptPOCUS and CCA being best detected in exit 3. The table also shows that FPR80 is lower than FPR95 for all exits and OOD data sets.

To visualise the separate energy scores from the different exits, the distribution for each OOD data set is shown in Fig. 6. Corresponding plots for the uncertainties from the deep ensemble can be seen in Fig. 7. For all methods and OOD data sets there is an overlap of the distribution for the POCUS test set and OOD data sets, making it impossible to separate them perfectly.

Finally, all methods were evaluated for breast cancer classification on the POCUS test set, see Table 5. Since the underlying classification networks are trained differently for the different methods, the baseline results on the whole test set (without leaving out non-trustworthy samples) varies. With a leave-out-rate of 0%, the deep ensemble performs the best, with an AUC on 95.6%. When including thresholds for detecting OOD samples, all methods except for the multi-level energy score show an improved classification performance on the remaining samples. Using a 5% leave-out-rate, the highest performance was obtained when using an ensemble with weighted variance-based uncertainty, with an AUC of 96.1%. The best performance for the 20% leave-out-rate was achieved using an ensemble with variance-based uncertainty, with an AUC of 97.6%.

Table 3. AUC (%) and FPR (%) for the different OOD detection methods evaluated on the OOD data sets. FPR is measured at TPR 95% and 80% for the POCUS training data (FPR95 and FPR80). Here ↓ implies smaller values are superior and ↑ implies larger values are superior.

Method	OOD data	AUC ↑	AUC 95% CI ↑	FPR95 ↓	FPR80 ↓
Softmax	MNIST	0.1	0.0–0.3	100.0	100.0
	CorruptPOCUS	73.1	70.0–76.3	96.0	63.7
	CCA	53.6	46.6–61.7	100.0	69.0
Energy	MNIST	98.1	97.3–98.9	1.0	0.0
	CorruptPOCUS	85.5	82.9–88.2	25.8	18.6
	CCA	78.5	72.5–84.0	60.7	31.0
Ensemble with variance	MNIST	79.1	77.1–80.9	56.0	41.7
	CorruptPOCUS	97.9	97.2–98.5	7.7	4.1
	CCA	70.1	65.0–74.4	88.1	71.4
Ensemble with weighted variance	MNIST	80.4	78.6–82.1	52.8	37.4
	CorruptPOCUS	97.2	96.2–98.0	8.7	6.2
	CCA	71.2	66.2–75.5	83.3	67.9
Ensemble with entropy, epistemic uncertainty	MNIST	84.6	82.8–86.3	48.6	27.7
	CorruptPOCUS	97.4	96.3–98.2	7.3	5.1
	CCA	70.2	65.4–74.6	85.7	69.0

Table 4. OOD detection results for multi-level energy score measured in AUC (%) and FPR (%) for each exit and OOD data set. FPR is measured at TPR 95% and 80% for the POCUS training data (FPR95 and FPR80). Here ↓ implies smaller values are superior and ↑ implies larger values are superior.

OOD data	AUC ↑			FPR95 ↓			FPR80 ↓		
	exit			exit			exit		
	1	2	3	1	2	3	1	2	3
MNIST	29	98	99	99	7	0	95	2	0
CorruptPOCUS	83	73	88	24	36	21	20	32	16
CCA	20	67	87	100	58	45	96	45	29

Using softmax score lead to a significant improvement in classification using a leave-out-rate of both 5% and 20%, meaning the remaining predictions are more likely to be correct and hence more trustworthy. All deep ensemble methods also show significant improvements, however the epistemic uncertainty only shows significant improvement using the 20% leave-out-rate and no significant improvement for the 5% leave-out-rate. The variance-based and weighted variance-based ensemble methods show significant improvements for both leave-out-rates.

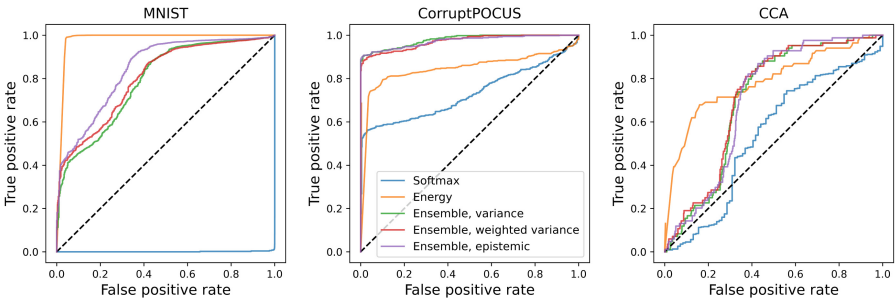


Fig. 5. ROC curves for the OOD detection methods evaluated on MNIST (left), CorruptPOCUS (middle) and CCA (right).

6 Discussion

The softmax score achieved inferior OOD detection performance compared to the other methods on all data sets, with the lowest AUC and highest FPR for all OOD data sets, see Table 3 and Fig. 5. It has previously been shown that the softmax score can be overconfident for samples far away from the training data [18], which our results corroborate. The energy scores from different exits are useful for different OOD data, as can be seen in Table 4. The MNIST samples are not well detected as OOD in the first exit, but in the last two exits they are almost perfectly detected as OOD. This is supported in Fig. 6 where at the last exit the energy distribution for MNIST is almost totally separated from the distribution for the POCUS test set. The CorruptPOCUS and CCA data are best detected in the third exit, but are not as easily separated from the ID data as the MNIST data. Table 4 also displays that the threshold at FPR80 achieves better results than at FPR95, specifically for CCA at exit 3 where FPR80 is 29% compared to 45% for FPR95. This general pattern is also confirmed in Table 3 where all methods achieve a better FPR at the 80% threshold compared to the 95% threshold.

According to Table 3, all OOD detection methods struggle to detect the CCA data. These images are ultrasound images similar to the ID data, with the difference that they are capturing the common carotid artery instead of breast tissue, hence they represent the second case mentioned in Sect. 1. Since these images are so similar to the ID data, the OOD detection methods have trouble separating them from the POCUS test data.

In a real world setting, which data is OOD heavily depends on the training data, the training procedure, hyperparameters and the network itself. What appears to the human eye as OOD might not be OOD from a computational perspective and vice versa. Underlying structures in an image or noise can cause an image to be OOD without the human eye noticing these difficulties for the algorithm. Therefore, there might be more cases of OOD data then the three cases discussed here. In an optimal setting, an OOD detection algorithm would

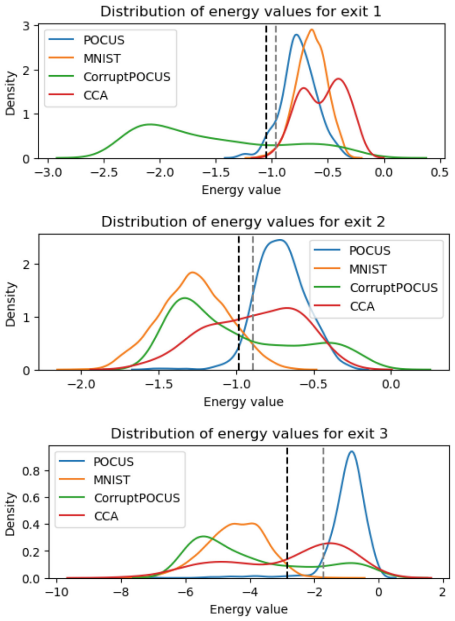


Fig. 6. Distribution of energy scores for the OOD data sets. Energies from exit 1 (left), exit 2 (middle) and exit 3 (right). The vertical dashed lines mark the threshold where 80% (light gray) and 95% (black) of the POCUS training set images are detected as ID.

show a strong correlation between OOD detection scores and the correctness of a prediction, covering not only the cases discussed here, but also images that are OOD due to other reasons. Ideally, we therefore want to find an OOD score threshold that excludes all kind of unsuitable cases, including noisy images, images of wrong tissue, images of rare lesions, but also images where the algorithm is very uncertain on how to make a trustworthy prediction. A well working OOD detection method would therefore lead to improved predictive accuracies when data samples with high OOD scores are excluded.

Removing OOD samples with softmax score does not have any larger impact on the AUC for the classification, see Table 5. Using the multi-level energy score for removing OOD samples decreases the AUC, implying that there is no correlation between correctness of prediction and energy score. However, the performance of the classifier improves significantly when removing OOD samples with the uncertainties from the deep ensemble. Additionally, the deep ensemble methods perform superior when detecting CorruptPOCUS as OOD, which is corrupted ID data (case 1 in Sect. 1). The good performance for both OOD detection and classification implies that using the uncertainties works well when the OOD data is relatively similar to the ID data. The deep ensemble methods and multi-level energy score perform well on detecting different types of OOD

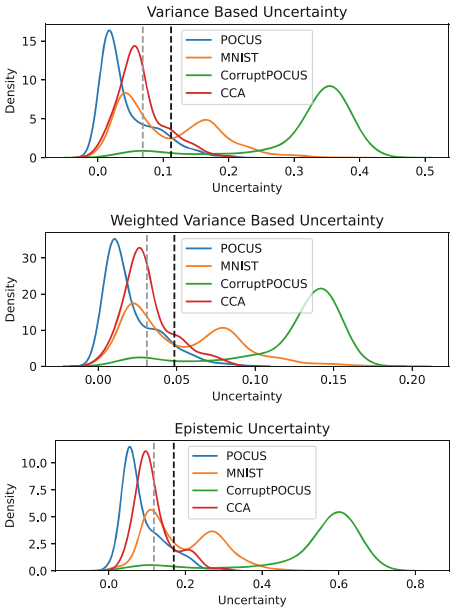


Fig. 7. Distribution of the uncertainties from the deep ensemble methods for each OOD data set: variance-based uncertainty (top), weighted variance-based uncertainty (middle) and entropy-based epistemic uncertainty (bottom). The vertical dashed lines mark the threshold where 80% (light gray) and 95% (black) of the POCUS training set images are detected as ID.

data. A combination of these two OOD detection methods might be a good solution to detect more cases and increase the trustworthiness of the classifier.

The softmax score and multi-level energy score have the advantages of not requiring multiple trainings of the network, hence are fast to use. Deep ensembles come with the drawback of complexity, in our case having a 20 times higher training and inference time compared to the other methods.

Even though the OOD detection methods might not work as well for data very close to the ID data compared to data that is very clearly OOD, it shows promising results detecting data capturing corrupted ultrasound images. In a real world setting this has the potential of being useful, by having the OOD detector flag when an image is unsuitable and the prediction should not be trusted. The novelty of this research lies in the application for a POCUS based automatic breast cancer classification, which only becomes applicable in the real world when the safety of the algorithm can be ensured. For future research more OOD detection methods should be investigated, for example Bayesian neural networks, deterministic uncertainty quantification methods and post-hoc OOD detection methods. Furthermore, the generalizablity of these methods across different medical applications should be tested.

Table 5. The classification results (cancer versus non-cancer) for each OOD detection method, measured with AUC (%) and AUC 95% CI. Here ↓ implies smaller values are superior and ↑ implies larger values are superior.

Method	Leave-out-rate (%)	AUC ↑	95% CI AUC ↑
Softmax	0	94.2	92.3–96.0
	5	95.2	93.2–96.9
	20	95.4	92.9–97.4
Energy	0	94.6	92.4–96.2
	5	94.2	92.1–96.2
	20	93.3	90.8–95.5
Ensemble with variance	0	95.6	93.8–97.0
	5	95.9	94.1–97.4
	20	97.6	95.7–99.1
Ensemble with weighted variance	0	95.6	93.8–97.0
	5	96.1	94.4–97.7
	20	97.4	95.5–98.9
Ensemble with entropy, epistemic uncertainty	0	95.6	93.8–97.0
	5	95.5	93.5–97.2
	20	97.3	95.3–98.7

7 Conclusion

In this work, three different OOD detection methods have been compared and evaluated for the novel application of breast cancer classification in POCUS. Due to different types of possible OOD cases, all methods were evaluated on three different OOD data set before integrating them into the breast cancer classification pipeline. The multi-level energy score performed the best on the MNIST and CCA data, while the deep ensembles were superior on the CorruptPOCUS data. Since the different OOD detection methods have proven to perform well on different OOD data, it could be promising to investigate the use of different OOD detection methods combined. The relative complexity of the deep ensemble requires more computational power compared to using the multi-level energy score. Thus, there is a balance between performance and computational complexity when it comes to OOD detection. Finding OOD samples is important if deep learning algorithms are to be used safely in a real world medical setting. The methods show promising result for detecting OOD data far from the ID data, but further research is needed in order to detect OOD data very similar to the ID data, including rare types of cancer and cases with atypical appearances.

Acknowledgement. This work was supported by strategic research area eSSENCE and Analytic Imaging Diagnostics Arena (AIDA), Vinnova Grant 2021-01420.

Compliance with Ethical Standards. This study was performed in line with the principles of the Declaration of Helsinki. Approval was granted by the Swedish Ethical Review Authority of Region Skåne (2019-04607).

References

1. Kevin Zhou, S., et al.: A review of deep learning in medical imaging: imaging traits, technology trends, case studies with progress highlights, and future promises. Proc. IEEE **109**(5), 820–838 (2021)
2. Yang, J., Zhou, K., Li, Y., Liu, Z.: Generalized out-of-distribution detection: a survey. CoRR, vol. abs/2110.11334 (2021)
3. Nemani, V., et al.: Uncertainty quantification in machine learning for engineering design and health prognostics: a tutorial. Mech. Syst. Signal Process. **205**, 110796 (2023)
4. Sung, H., et al.: Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J. Clin. **71**(3), 209–249 (2021)
5. Francies, F.Z., Hull, R., Khanyile, R., Dlamini, Z.: Breast cancer in low-middle income countries: abnormality in splicing and lack of targeted treatment options. Am. J. Cancer Res. **10**(5), 1568 (2020)
6. World Health Organization. Global breast cancer initiative implementation framework: assessing, strengthening and scaling up of services for the early detection and management of breast cancer (2023)
7. Karlsson, J., et al.: Classification of point-of-care ultrasound in breast imaging using deep learning. In: Medical Imaging 2023: Computer-Aided Diagnosis. International Society for Optics and Photonics, vol. 12465, p. 124650Y. SPIE (2023)
8. Hendrycks, D., Gimpel, K.: A baseline for detecting misclassified and out-of-distribution examples in neural networks. arXiv preprint arXiv:1610.02136 (2016)
9. Liu, W., Wang, X., Owens, J., Li, Y.: Energy-based out-of-distribution detection. In: Advances in Neural Information Processing Systems (2020)
10. Lin, Z., Roy, S.D., Li, Y.: MOOD: multi-level out-of-distribution detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2021)
11. Hansen, L.K., Salamon, P.: Neural network ensembles. IEEE Trans. Pattern Anal. Mach. Intell. **12**(10), 993–1001 (1990)
12. Lakshminarayanan, B., Pritzel, A., Blundell, C.: Simple and scalable predictive uncertainty estimation using deep ensembles. In: Advances in Neural Information Processing Systems, vol. 30 (2017)
13. Hüllermeier, E., Waegeman, W.: Aleatoric and epistemic uncertainty in machine learning: an introduction to concepts and methods. Mach. Learn. **110**(3), 457–506 (2021)
14. GE Healthcare. Vscan Air. https://vscan.rocks/products-and-solutions/vscan-air-cl. Accessed 18 Mar 2024
15. Deng, L.: The MNIST database of handwritten digit images for machine learning research. IEEE Signal Process. Mag. **29**(6), 141–142 (2012)
16. Zukal, M., Beneš, R., Číka, P., Říha, K.: Ultrasound image database. http://splab.cz/en/download/databaze/ultrasound. Accessed 08 Nov 2023
17. LeNail, A.: NN-SVG: publication-ready neural network architecture schematics. J. Open Sour. Softw. **4**(33), 747 (2019)
18. Nguyen, A.M., Yosinski, J., Clune, J.: Deep neural networks are easily fooled: high confidence predictions for unrecognizable images. CoRR, vol. abs/1412.1897 (2014)

Colon Segmentation Using Guided Sequential Episodic Training and Contrastive Learning

Samir Harb[1,2(✉)], Asem Ali[1], Mohamed Yousuf[1,3], Salwa Elshazly[4], and Aly Farag[1]

[1] Computer Vision and Image Processing Laboratory (CVIP), University of Louisville, Louisville, KY, USA
safara01@louisville.edu
[2] Higher Technological Institute, 10th of Ramadan City, Egypt
[3] Faculty of Engineering, Ain Shams University, Cairo, Egypt
[4] Kentucky Imaging Technologies, Louisville, KY, USA

Abstract. Accurate colon segmentation on abdominal CT scans is crucial for various clinical applications. In this work, we propose an accurate approach to colon segmentation from abdomen CT scans. Our architecture incorporates 3D contextual information via sequential episodic training (SET). In each episode, we used two consecutive slices, in a CT scan, as support and query samples in addition to other slices that did not include colon regions as negative samples. Choosing consecutive slices is a proper assumption for support and query samples, as the anatomy of the body does not have abrupt changes. Unlike traditional few-shot segmentation (FSS) approaches, we use the episodic training strategy in a supervised manner. In addition, to improve the discriminability of the learned features of the model, an embedding space is developed using contrastive learning. To guide the contrastive learning process, we use an initial labeling that is generated by a Markov random field (MRF)-based approach. Finally, in the inference phase, we first detect the rectum, which can be accurately extracted using the MRF-based approach, and then apply the SET on the remaining slices. Experiments on our private dataset of 98 CT scans and a public dataset of 30 CT scans illustrate that the proposed FSS model achieves a remarkable validation dice coefficient (DC) of 97.3% (Jaccard index, JD 94. 5%) compared to the classical FSS approaches 82.1% (JD 70.3%). Our findings highlight the efficacy of sequential episodic training in accurate 3D medical imaging segmentation. The codes for the proposed models are available at https://github.com/Samir-Farag/ICPR2024.

Keywords: Colon segmentation · Deep learning · Few-shot

This work has been funded by NSF grant 2124316.

A. Antonacopoulos et al. (Eds.): ICPR 2024, LNCS 15313, pp. 64–79, 2025.
https://doi.org/10.1007/978-3-031-78201-5_5

1 Introduction

Automated image segmentation plays a crucial role in medical imaging research and clinical applications by automating or facilitating the delineation of anatomical structures and other regions of interest. The segmentation step is of significant importance to facilitate accurate identification and delineation of structures or abnormalities for clinical applications such as lesion localization, disease diagnosis, and prognosis [30,31]. Specifically, automatic colon segmentation is a key step for medical image analysis pipelines (e.g. colonography [3,16]), because any inaccuracies at the segmentation stage will carry through to subsequent steps. This underscores the importance of prioritizing the segmentation process and improving its effectiveness, which consequently leads to performance enhancements in the next stages of this pipeline.

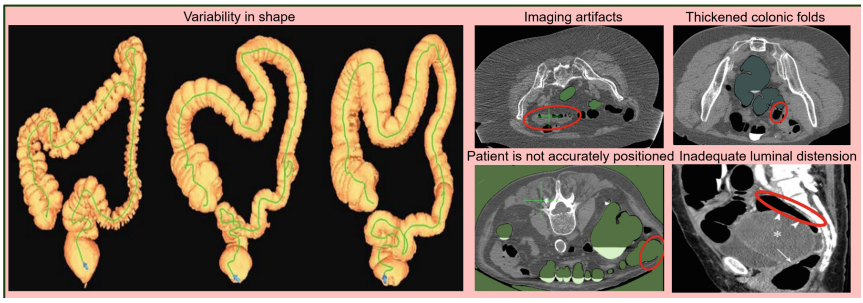


Fig. 1. Examples of challenges that hinder accurate segmentation of the colon [3,11].

However, segmenting the colon regions accurately from abdominal CT scans poses significant challenges, as depicted in Fig. 1. First, colon regions exhibit highly variable and asymmetric topology [22], and their positions vary between different CT images [9]. Second, distinguishing colon regions from surrounding structures is complicated by the presence of Hounsfield intensity regions containing soft tissues, air regions resembling gas-filled organs like the small intestine, and high-attenuation structures, e.g., bones. Lastly, patient preparation imperfections, such as residual stool and lesions, can lead to disjointed colon segments. These complexities inherent in colon segmentation, particularly in scenarios where automated algorithms are indispensable, may confuse segmentation algorithms [9]

Colon segmentation approaches that have been reported in the literature could be grouped into two main categories: (1) classic segmentation approaches, which typically employ techniques such as MRF-based models, e.g. [3,5,21,26], edge detection, region growth and division, e.g. [6,7,17,20,21], or hybrid segmentation algorithms [14]; and (2) deep learning (DL) approaches, e.g. [2,10,15,30], which exploit the available data to learn complicated high-level characteristics

that can be used for segmentation, unlike the classical approaches that focus on low-level traits, which may not be as helpful for segmentation.

Although DL approaches are successful segmentation tools, Jakob Wasserthal et al. [30], who developed the state-of-the-art (SOTA) segmentation tool, Totalsegmentator, reported that the colon posed the most significant challenges, with a failure rate of ∼35% of cases. This failure was mainly attributed to difficulties in precisely segmenting the subtle details of the colon.

Traditional deep Convolutional Neural Networks (CNNs) adept at semantic segmentation often encounter challenges, relying on a plethora of densely annotated images for effective training and struggling to generalize to unfamiliar object classes. This issue is exacerbated in medical imaging, where the dearth of annotations hampers the applicability of conventional methods. Recently, few-shot learning (FSL) has emerged as a prominent deep learning approach to equip a model with the ability to segment unseen semantic classes by learning from just a few labeled images of this unseen class during inference, without necessitating model retraining. Hence, Few-Shot Segmentation (FSS) was introduced to address the challenges of medical image segmentation by leveraging knowledge distilled from labeled samples (support) to segment unlabeled samples (query). FSS learns tasks composed of base class in an episodic training manner and segments unseen classes in the form of tasks in the inference stage.

One of the pioneering DL networks proposed to utilize FSL in natural images is PANet [29]. Prior to PANet, FSS methods demonstrated unsatisfactory generalization due to a lack of separation between knowledge extraction and segmentation processes, as well as the utilization of support data solely for masking purposes. PANet addressed these issues by introducing a separation between prototype extraction (which involves feature extraction from support images and subsequent prototype extraction from these features, along with feature extraction from query images) and non-parametric metric learning (which segments the query image by computing the cosine distance between each support class prototype and query features at each spatial location). Furthermore, PANet uses annotations to supervise Few-Shot Learning. To eliminate the need for annotations during training, Ouyang, Cheng et al. [19] developed a self-supervised FSS framework, SSL-ALPNet, that exclusively utilizes superpixel-based pseudo-labels for supervision. In addition, an adaptive local prototype module is presented to mitigate the challenge of foreground-background imbalance in medical image segmentation. Wu, Huisi et al. [31] proposed AAS-DCL to combine dual contrastive learning and anatomical guidance to enhance feature discriminability and data utilization to help few-shot medical image segmentation.

In this work, we propose a novel FSS approach for precise colon segmentation in abdominal CT scans, addressing the inherent challenges of this critical medical imaging task. Our proposed approach introduces an episodic segmentation strategy that takes advantage of sequential episode training and contrastive learning techniques. Unlike traditional few-shot segmentation approaches, our method employs supervised episodic training, facilitating enhanced feature discriminability and segmentation accuracy. In particular, we incorporate unrelated

slices rich in anatomical structures to provide vital background guidance, further refining the segmentation process. Based on the AAS-DCL framework [31], our approach integrates dual contrastive learning (DCL) and anatomical guidance, culminating in improved feature extraction and segmentation performance. In addition, we introduce a novel MRF-based rectum detection and initial labeling technique, enhancing the robustness and accuracy of the proposed approach. The primary contributions of our work are as follows:

i) Develop an MRF-based rectum detection and initial labeling method, contributing to improved accuracy and robustness of the overall segmentation process.
ii) Integrate supervised sequential episodic training and contrastive learning techniques to enhance feature discriminability and segmentation accuracy, while incorporating unrelated slices rich in anatomical structures to provide essential background guidance.
iii) Enhance feature extraction and segmentation performance through the integration of dual contrastive learning with anatomical guidance.

2 Method

Our approach aims to accurately segment the colon in abdominal CT scans. We use a method that combines 3D information (through SET) with 2D segmentation models. This allows us to avoid the high computational costs of complex 3D neural networks while still achieving precise results. The 2D models are efficient and flexible, handling individual CT images well even with irregular sampling.

2.1 Proposed Episodic Segmentation Approach

The traditional episodic training strategy for the few-shot segmentation (FSS) approach involves training a model over a large number of epochs, with multiple episodes in each epoch. So, a dataset, in episodic training, is arranged into multiple episodes and each episode consists of support and query pairs. For a set of images $\mathcal{X}$ and its corresponding set of binary masks $\mathcal{Y}$, we define the support set $\mathcal{S} = \{x_s^c, y_s^c\}$ and the query set $\mathcal{Q} = \{x_q^c, y_q^c\}$, where $x_{s(q)}^c \in \mathcal{X}$, $y_{s(q)}^c \in \mathcal{Y}$, and the superscript c represents an arbitrary class in a set of classes $\mathcal{C}$. Since few-shot segmentation approaches were introduced to take advantage of distilled knowledge from labeled samples for segmenting unlabeled ones, in these approaches, a model is trained to identify a set of classes $\mathcal{C}_{tr}$ in a training dataset $\mathcal{D}_{tr}$. But it never sees the set of classes $\mathcal{C}_{ts}$ in the test dataset $\mathcal{D}_{ts}$. Then, during the inference, the model is used to segment the unseen classes $\mathcal{C}_{ts}$ in $\mathcal{D}_{ts}$ using annotated samples of these classes, without the need to re-train the model.

We propose an FSS-like approach in which we use support and query sets, but unlike the classical FSS approaches, we use the episodic training strategy in a supervised manner. Therefore, training and test classes are the same, that is, $\mathcal{C}_{tr} = \mathcal{C}_{ts} = \{colon\}$, but $\mathcal{D}_{tr} \neq \mathcal{D}_{ts}$ where $\mathcal{D}_{tr}$ contains training scans and

$\mathcal{D}_{ts}$ contains testing scans. Moreover, to enhance the discriminability of learned features of the model, we exploit unrelated slices $U = \{x_u^{\bar{c}}, y_u^{\bar{c}}\}$ but rich with anatomical structures (i.e., $\bar{c}$ non-colon regions, e.g., liver) to provide more background guidance. Using support, query, and unrelated features, extracted from x_s^c, x_q^c, and $x_u^{\bar{c}}$, respectively, we develop an embedding space using contrastive learning to pull closer (x_s^c, x_q^c) and push farther $(x_q^c, x_u^{\bar{c}})$.

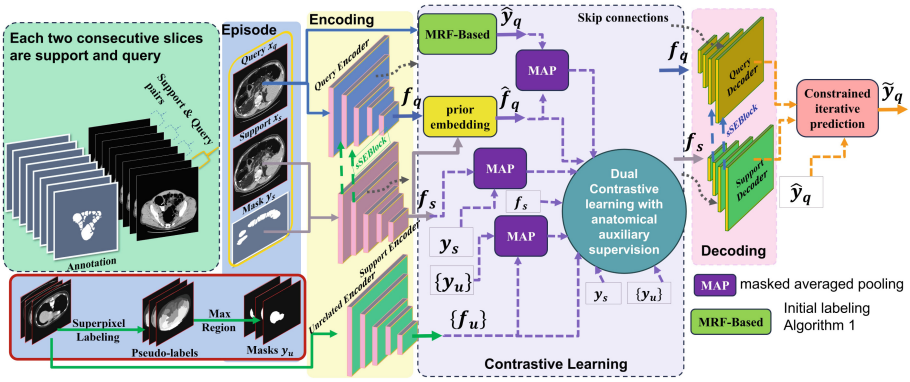


Fig. 2. The proposed SET FSS approach with DCL. An MRF-based auxiliary supervision is used to enhance the baseline AAS-DCL. The workflow starts by arranging scans into pairs of consecutive slices (x_s^c, x_q^c). Then unrelated slices are selected as negative samples $\{x_u^{\bar{c}}\}$. The episode $(x_s^c, x_q^c, \{x_u^{\bar{c}}\})$ is fed into encoders to extract features $(f_s, f_q, \{f_u\})$. Masked average pooling is applied on these features and contrastive learning is used to generate an embedding space. Finally, decoders with skip connection are used to estimate the final segmentation $\widetilde{y}_q$, which is iteratively refined using initial labeling.

We build on the AAS-DCL approach [31], as shown in Fig. 2, to combine DCL and anatomical guidance to enhance feature discriminability. However, unlike the AAS-DCL approach, we use SET, in which support and query samples are consecutive slices in a CT scan. The motivation behind this is that the anatomy of the human body does not have abrupt changes and thus if a pixel in the current CT slice is colon, it is most likely that this pixel will be colon within the next or the previous few CT slices. Therefore, using consecutive slices as support and query samples simplifies the segmentation approach. Moreover, for more guidance, we start with an initial labeling that is generated using an MRF-based approach (Algorithm 1). This enhances the DCL. Finally, in the inference stage, we first detect the rectum, which can be accurately extracted as shown in Fig. 4, then apply the sequential episodic approach.

Unrelated Slices Selection: The proposed workflow starts by arranging CT scans into pairs of consecutive slices (x_s^c, x_q^c). Then, in each episode, we randomly

select three unrelated slices as negative samples $\{x_u^{\bar{c}}\}$. These unrelated slices do not include colon regions but they may have irrelevant organs or tissues. To define masks $\{y_u^{\bar{c}}\}$ for unrelated samples $\{x_u^{\bar{c}}\}$, we employ an unsupervised graph cut-based algorithm [8], offline, which generates superpixel segmentation. These pseudo-labels are binarized by choosing the dominant superpixel (i.e., the largest connected region) in each pseudo-label as a target and other superpixels as a background. Then an encoder is used to extract the features $\{f_u^{\bar{c}}\}$ from $\{x_u^{\bar{c}}\}$. Finally, these features and their masks $\{f_u^{\bar{c}}, y_u^{\bar{c}}\}$ are included in the AAS-DCL scheme.

Dual Contrastive Learning: To provide more background guidance, we exploit the unrelated slices with query and support slices in contrastive learning. Inspired by the baseline AAS-DCL approach [31], we combine prototypical contrastive learning and contextual contrastive learning to form a DCL scheme, which makes the features of the colon regions closer to other characteristics of dissimilar tissues. The infoNCE loss [18] $\mathcal{L}(v_q, v_s, v_u)$ is used for the training process of the contrastive learning module.

$$\mathcal{L}(v_q, v_s, v_u) = -v_q.v_s/\tau + \log \sum_{i=1}^{n} \exp(v_q.v_{ui}/\tau),$$

where τ is a control parameter, n is the number of negative samples, v_q, v_s, and v_u are the query, support and background prototypes, respectively. These prototypes are generated by the global average pooling of features and the corresponding masks. However, these prototypes cannot acquire intra-class variations. To overcome this problem, patch-based prototypes may be used.

Prototypical Contrastive Learning: Prototype-based learning is based on the generation of prototypes that discriminate between the features of the foreground and the background. In this approach, support features $\{f_s\}$ and their corresponding masks $\{y_s\}$ are used to generate the colon prototype using the masked averaged pooling (MAP) operation [34] $v_s = \frac{\sum_r y_s(r).f_s(r)}{\sum_r y_s(r)}$.

Unlike the baseline AAS-DCL approach, which uses global average pooling to calculate the query prototype, we exploit the initial query mask $\hat{y}_q$ to calculate the query prototype using masked average pooling. Also, instead of using the query feature f_q, we employ a prior embedding module [31] to enhance the query feature. The enhanced query feature $\hat{f}_q$ further activates the foreground information in the query prototype $v_q = \frac{\sum_r \hat{y}_q(r).\hat{f}_q(r)}{\sum_r \hat{y}_q(r)}$.

Similarly, unrelated features $\{f_u\}$ and their corresponding masks $\{y_u\}$ are also used to generate the background prototype v_u using masked averaged pooling.

To overcome intra-class variations and to exploit information about other structures around colon regions as unrelated samples, we employed patch-based learning [19]. In this method, the support feature and its mask are divided into

patches, then these patches are used to generate a colon prototype and a background prototype depending on a threshold. This scheme increases the number of negative samples and distinguishes between the local characteristics of different tissues. Again, unlike the baseline AAS-DCL approach, we exploit the initial mask of the query $\hat{y}_q$ to calculate the query prototype using the masked average pooling.

Contextual Contrastive Learning: Finally, to guide feature maps focusing on rich contextual information, a spatial attention block [24] is employed to process the support feature f_s, enhanced query feature $\hat{f}_q$ and unrelated features $\{f_u\}$. Then the processed features are averaged and used in contextual contrastive learning, for more details see [31].

Iterative Prediction: For accurate segmentation, iterative optimization methods [28,32] are used to combine the prediction of the query with the query feature by convolution. Unlike the baseline [31], we guide the iterative process using the initial labeling to promote the fusion of the query feature and the predicted mask. The query prediction is updated through the similarity consistency constraint, in which we also use initial labeling to calculate a similarity map between support and query features.

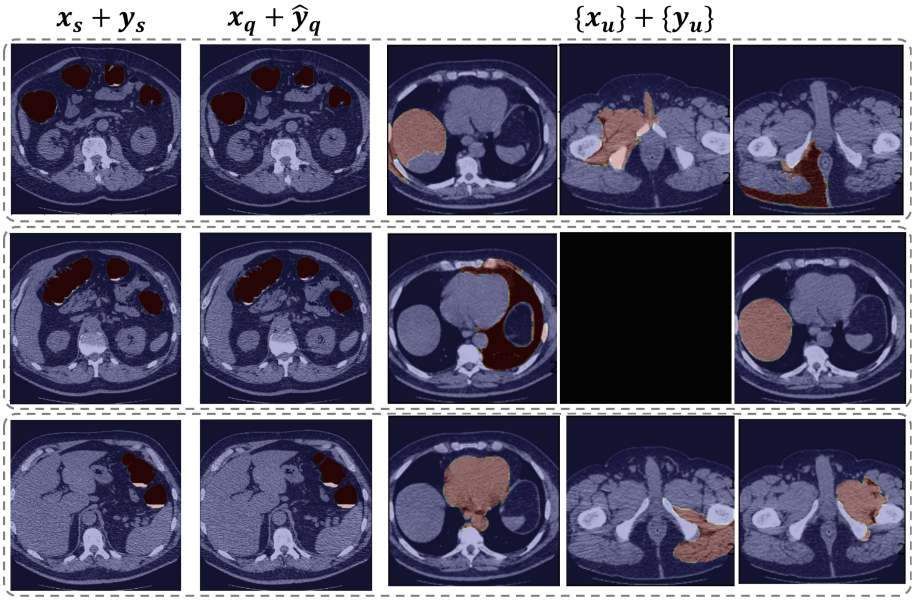


Fig. 3. Examples of episodes from training dataset. Each row represents a single episode that includes labeled support, query with initial labeling, and unrelated labeled slices.

Training Stage: In this stage, two consecutive slices are randomly selected as a pair of support and query. In addition, three randomly selected unrelated slices are added to this pair to form an episode, as shown in Fig. 3. Each episode is fed into the encoder-decoder sSENet [25] for feature extraction and reconstruction. Then the cross-entropy loss uses the prediction of this module to calculate a prediction error against the ground truth. The prediction error and contrastive learning loss, which are computed using the extracted features and the initial query mask, are backpropagated to train the network.

Inference Stage: This stage starts by detecting the rectum and the initial mask for a given abdomen CT scan using the proposed MRF-based approach. Subsequently, a rectum slice is considered a support sample and the consecutive slice is a query. In addition, three randomly selected unrelated slices are added to this pair to form an episode. Each episode is fed into the trained model to generate the prediction of the query. Then, the segmented query slice will be the support sample for the consecutive slice in the sequence. This iterative process continues until all colon regions are successfully segmented.

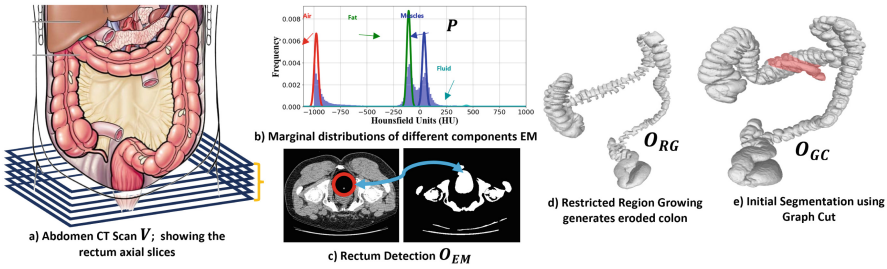


Fig. 4. MRF-based initial labeling approach. Rectum is the only region, in the lower CT slices that has air, and it can be easily identified as a disk-like region that has low Hounsfield. First, EM is used to calculate the empirical distributions **P** of Hounsfield intensities in a DICOM volume **V**. Thresholds between air and fat and between muscle and fluid are used to generate $\mathbf{O}_{EM}$. RG algorithm is applied starting from the rectum to generate an eroded colon $\mathbf{O}_{RG}$. This guarantees that other organs, e.g., small intestine, are not merged with $\mathbf{O}_{RG}$. Finally, $\mathbf{O}_{RG}$ is refined through an optimization technique to generate $\mathbf{O}_{GC}$, which still may have other structures (colored) misclassified as colon. (Color figure online)

2.2 MRF-Based Rectum Detection and Initial Labeling

To generate an initial labeling, we develop a multi-step approach, which employs three algorithms: Expectation-Maximization (EM) to calculate the empirical distributions of Hounsfield units (HU) in a DICOM volume, Region Growing

(RG) to generate initial labeling by identifying colon regions starting from the rectum, and Graph Cut (GC) to estimate the initial mask of colon regions. The main components of a colon are the air, for which the characteristic peaks are almost at -1000 HU [17], and the opacified fluid whose Hounsfield intensity is greater than 300 HU. To extract the colon components, first we estimate the marginal densities of air, fat, muscle, and fluid, in an abdomen scan, by fitting four Gaussian components using the EM algorithm, as shown in Fig. 4-b. Then, we identify colon regions using two thresholds. As shown in Fig. 4-c, the rectum is the only region, in the lower CT slices of an abdomen scan, that has air. Therefore, the rectum region can be easily identified as a disk-like region with a low Hounsfield unit. This region is used as a starting seed, from which other colon regions are extracted by the proposed model.

The problem is formulated as the maximum-A posterior estimate of an MRF model, which involves finding the labeling that minimizes the following energy function $E(\tilde{y})$ (Eq. (1)) that combines both the spatial smoothness and data consistency.

$$E(\tilde{y}) = \sum_{\{r,t\}\in\mathcal{N}} V(\tilde{y}_r, \tilde{y}_t) + \sum_{r\in\mathcal{P}} D(\tilde{y}_r), \tag{1}$$

where $\mathcal{N}$ represents the set of neighboring pixel pairs $(r, t), V(.,.)$ is the potential function that penalizes label inconsistencies between neighboring pixels, and $D(.)$ is the data penalty term that measures how well the labeling $\tilde{y}_r$ matches the observed data. The minimization of the energy function in Eq. (1) using a graph cut generates the initial labeling result. The Algorithm 1 summarizes this approach.

Algorithm 1. MRF-based segmentation approach

0: **Input**: DICOM volume **V**
1: Calculate the histogram **P** of HU values in **V**
2: Apply EM algorithm, and identify air and fluid regions O_{EM}
3: Detect rectum region in O_{EM}
4: Starting from rectum region, apply RG algorithm to extract colon O_{RG} from O_{EM}
5: Use O_{RG} as a seed for GC and minimize $E(\tilde{y})$ to extract initial labeling O_{GC}

3 Experiments

Dataset: We conducted experiments on our private dataset having abdominal CT scans of 49 patients in both supine and prone positions. Experts annotated the colon segments in these 98 CT scans. Also, for the sake of comparison, we use the synapse public dataset [12] which has been used by several SOTA approaches. In our work, we refer to this dataset as SABS. It contains 30 abdominal CT scans. In SABS dataset, 13 organs were manually annotated (colon is not included) by 2 experienced undergraduate students and verified by a radiologist [1]. From

these two datasets, we created three training datasets, $\mathcal{D}_{tr_1}$, $\mathcal{D}_{tr_2}$ and $\mathcal{D}_{tr_3}$, and a testing dataset $\mathcal{D}_{ts}$:

i $\mathcal{D}_{tr_1}$ **(SABS):** Consists of 30 scans from the SABS CT dataset, with annotated organs labeled 1 through 13. Specifically, the labels include spleen (1), right kidney (2), left kidney (3), and so forth, up to the left adrenal gland (13).
ii $\mathcal{D}_{tr_2}$ **(SABS + CTC68):** Combines the SABS dataset (with 13 annotated organs) and 68 scans (34 prone and 34 supine) from our private dataset (with annotated colon). Consequently, the combined dataset covers spleen (1), right kidney (2), left kidney (3), and so forth, up to the left adrenal gland (13), and includes the colon as label 14.
iii $\mathcal{D}_{tr_3}$ **(CTC68):** Includes 68 scans (34 prone and 34 supine) from our private dataset (with annotated colon).
iv $\mathcal{D}_{ts}$ **(CTC30):** Encompasses 30 scans (15 prone and 15 supine) from our private dataset (with annotated colon).

Evaluation Metrics. We employed both DC and JD to quantify the pixel-wise agreement between the predicted and ground truth segmentation [27]. This dual assessment approach considers both the overlap and spatial agreement between the predicted and ground truth colon regions.

Technical and Implementation Details: We implemented our framework with PyTorch, based on the official baseline implementation https://github.com/cvszusparkle/AAS-DCL_FSS, on a Nvidia TITAN RTX with 24 GB. Among the different available off-the-shelf fully convolutional networks, we utilized ResNet101 that guaranteed high spatial resolutions in feature maps. As a preprocessing step, we first resize the 2D slices to 256×256 resolution and divide data into 4 patches for prototypical contrastive learning. Our proposed SET starts with a learning rate of 10^{-4}, a batch size of 1, and applies polynomial decay. Adam optimization with power = 0.95 and weight decay = 10^{-7} is used over 100 epochs. Data augmentation includes random adjustments to sharpness and lightness. For high-resolution feature maps, a fully convolutional ResNet101 pre-trained on MS-COCO processes 256×256 images to $256 \times 32 \times 32$ maps. Training uses a Local Pooling Window of 4×4, reducing to 2×2 for inference. Training on a Nvidia TITAN RTX GPU takes 3 h, using 3 GB memory, on average for the proposed model.

Standard FSS Approaches vs the Proposed SET FSS Approach: Since the proposed approach uses the FSS concept of support and query sets, we compare its performance against standard FSS approaches. To highlight the high performance of our proposed SET FSS approach with respect to the standard FSS segmentation approaches, we evaluated the SSL-ALPNet [19] model and the AAS-DCL [31] network, which is our baseline, in the colon segmentation problem. The experimental results on the test set $\mathcal{D}_{ts}$, shown in Table 1, shed light on the performance of various model configurations in colon segmentation.

Table 1. Comparison of validation DC and JD on $\mathcal{D}_{ts}$ dataset for our proposed models against the SOTA models, with different training and initialization settings.

	Model	Training	Initialization	DC	JD
SOTA	SSL-ALPNet [19]	SABS	None	34.1%	21.0%
		SABS + CTC	None	81.7%	70.0%
		CTC	None	82.1%	70.3%
	AAS-DCL [31]	SABS	None	16.0%	8.8%
		SABS + CTC	None	65.5%	49.5%
		CTC	None	68.8%	53.2%
Proposed	Guided-AAS-DCL	SABS	Superpixel	61.0%	44.2%
		SABS + CTC	Superpixel+MRF	83.0%	71%
		CTC	MRF	96.3%	92.9%
	SET-DCL	CTC	None	96.8%	93.7%
	Guided-SET-DCL	**CTC**	**MRF**	**97.3%**	**94.5%**

First, we used the standard FSS technique, in which we train SSL-ALPNet and AAS-DCL models using $\mathcal{D}tr1$ (i.e., self-supervised learning by training networks with data that included superpixel results instead of annotations). As expected, standard FSS techniques do not perform well in this scenario. This is due to many reasons, such as uncertainties in the dataset (e.g., prep deficits, patient conditions, and scanner settings and errors). In addition, the learned embedding space of the prototypes of different organs in the $\mathcal{D}tr1$ dataset has different distributions than the colon prototype due to the characteristics of different tissues. Specifically, SSL-ALPNet trained in SABS achieved 34.1% DC and 21.0% JD, and AAS-DCL trained on $\mathcal{D}tr1$ achieved 16.0% DC and 8.8% JD. For learning a more general embedding space, in the second experiment, we included the colon in the training phase. So, we used $\mathcal{D}tr2$ to train the two models (i.e., supervised learning by training networks with data including the colon along with the other 13 organs). This drastically enhances the performance of the models. SSL-ALPNet trained on $\mathcal{D}tr2$ achieved 81.7% DC and 70.0% JD, while AAS-DCL trained on $\mathcal{D}tr2$ achieved 65.5% DC and 49.5% JD.

To explore the upper limit of the performance of the models, we used the purely supervised learning scheme by training the models using $\mathcal{D}tr3$. The SSL-ALPNet trained model provides decent performance, achieving 82.1% DC and 70.3% JD, because the SSL-ALPNet model ensures that each prototype exclusively represents a distinct part of the object-of-interest. This enables precise localization of colon structures by preserving intricate local details crucial to segmentation accuracy. However, the AAS-DCL model needs more guidance to enhance its performance, achieving only 68.8% DC and 53.2% JD.

Ablation Study: The proposed approach depends on the initial labeling and sequential episodic learning. Table 1 summarize effects of these components. In order to enhance the performance of the baseline model, we guide the DCL

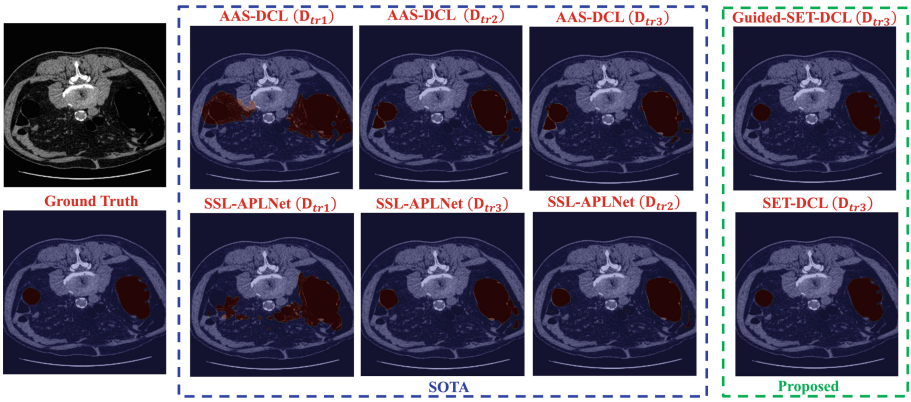


Fig. 5. Qualitative results on different training settings show that the results of SOTA FSS approaches include artifacts, on the other hand, the proposed method achieves desirable segmentation results that are close to ground truth.

using an initial labeling as explained in the proposed approach. Also, we add the constraint on a query slice to be within 5 neighbors from the support slice. This limits the changes in the colon structure. The first Guided-AAS-DCL model is trained using $\mathcal{D}tr1$ and the initial labeling for the organ of interest is estimated using the superpixel approach. The initialization and the neighbor constraint enhance the model performance from 16.0% to 61.0% DC and from 8.8% to 44.2% JD. Adding colon scans with MRF-based initial labeling to the training dataset in $\mathcal{D}tr2$ boosts the model performance, yielding a DC of 83.0% and a JD of 71%. Finally, the supervised learning performance of the Guided-AAS-DCL model reaches 96.3% DC and 92.9% JD. This highlights that the synergistic fusion of initial labeling and query constraint promises to deliver precise and reliable colon segmentation results.

Exploiting the anatomical structure of the colon, we propose the sequential episodic training SET-DCL FSS approach, in which the support and query are neighboring slices. Additionally, the inference phase starts with the detected rectum slices as a support and then sequentially segments the remaining slices where each segmented slice acts as a support slice for the consecutive query slice in the CT scan. Without any additional initialization, the proposed SET-DCL model exhibits a DC of 96.8% and a JD of 93.7%, better than the Guided-AAS-DCL model. Moreover, leveraging MRF-based initialization further enhances the performance of the proposed Guided-SET-DCL model's performance further, resulting in a remarkable DC of 97.3% and a JD of 94.5%. This underscores the efficacy of MRF-based initialization and sequential episodic training in increasing segmentation accuracy.

Figures 5 and 6 show the robustness of the proposed framework that consistently produces satisfactory results, especially for training solely with the CTC. Also, Fig. 7 provides more illustrations on how the proposed approach accurately

segments colon parts, while other SOTA approaches may miss parts and have some artifacts.

Supervised Learning Scheme: Finally, since our proposed approach depends on supervised learning, to compare against the SOTA CNN-based encoder-decoder segmentation architectures trained using a supervised learning scheme, we trained the PAN model [13] that is paired with resnest269e [33] backbone and U-Net model [23] using $\mathcal{D}tr3$ then we tested them on $\mathcal{D}_{ts}$. The primary challenge in traditional encoder-decoder networks lies in their inability to incorporate temporal information in a sequence of images such as colon CT scans. Therefore, we explore the fusion of C-LSTM with U-Net by replacing the convolutional layers in the encoder section with C-LSTM layers [4]. As shown in Table 2, our proposed approach outperforms the SOTA approaches. Specifically, the DC for our proposed approach (Guided-SET-DCL) is 97.3%, which is higher than MRF-based (87.9%), C-LSTMs (89.2%), U-Net (85.0%), and PAN (97.1%). Similarly, the JD for our proposed approach is 94.5%, which also outperforms MRF-based (84.5%), C-LSTMs (80.7%), U-Net (80.0%), and PAN (95.5%). The C-LSTM has a lower performance because it has a larger number of parameters that should be optimized, and this hinders the network learning, especially for long and high-resolution image sequences.

Although the experiments were conducted to segment the colon, we believe that the same approach can be successfully used to segment other organs that are scanned as sequential slices that do not have abrupt changes.

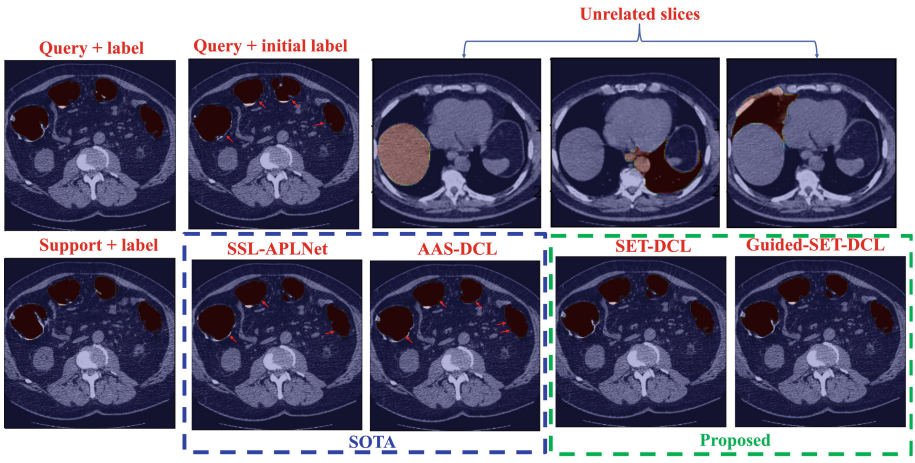


Fig. 6. An example of an episode: query with ground truth, query with initial labeling, unrelated slices with labels, and support with label. The qualitative results show that SOTA FSS approaches miss colon semilunar folds (shown in red arrows); on the other hand, the proposed method achieves desirable segmentation results that are close to ground truth. (Color figure online)

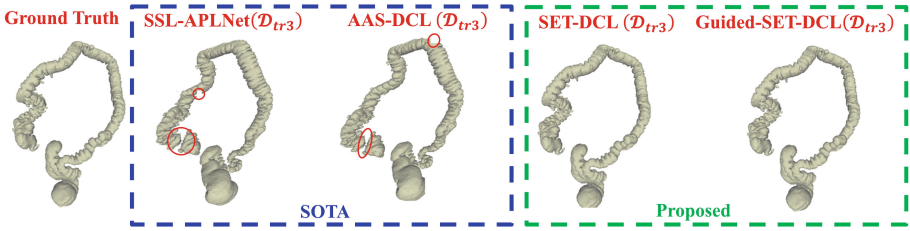


Fig. 7. Ground truth 3D colon and results of the proposed method compared to the SOTA FSS approaches. The qualitative results show that the SOTA FSS approaches miss parts and generate incomplete colon, on the other hand, the proposed method generates accurate colon segments.

Table 2. Comparison of validation DC and JD on $\mathcal{D}_{ts}$ dataset for our proposed approach against CNN-based SOTA architectures.

Metric	MRF-based	C-LSTMs [4]	U-Net [23]	PAN [13]	Guided-SET-DCL
DC	87.9%	89.2%	85.0%	97.1%	**97.3%**
JD	84.5%	80.7%	80.0%	**95.5%**	94.5%

4 Conclusions

We proposed an FSS approach that addresses the significant challenge of accurate colon segmentation in abdominal CT scans. Through the integration of a classical segmentation model, i.e., MRF model, deep learning, and sequential episodic training, we developed a comprehensive approach for colon segmentation. Using episodic training and dual contrastive learning, our Guided-SET-DCL approach achieves remarkable segmentation accuracy, outperforming traditional SOTA FSS methods and CNN-based models. We demonstrated the efficacy of our proposed approach in different training settings that highlighted its robustness and generalization capability. By incorporating sequential episodic training and anatomical guidance, we navigated the complexities of colon segmentation, overcoming challenges such as variable topology and variations in tissue intensity.

References

1. Multi-atlas labeling beyond the cranial vault - workshop and challenge. https://doi.org/10.7303/syn3193805. Accessed 3 Apr 2024
2. Akilandeswari, A., et al.: Automatic detection and segmentation of colorectal cancer with deep residual convolutional neural network. Evid.-Based Complement. Altern. Med. (2022)
3. Alkabbany, I., Ali, A.M., Mohamed, M., Elshazly, S.M., Farag, A.: An AI-based colonic polyp classifier for colorectal cancer screening using low-dose abdominal CT. Sensors **22**(24), 9761 (2022)

4. Arbelle, A., Raviv, T.R.: Microscopy cell segmentation via convolutional LSTM networks. In: 2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019), pp. 1008–1012. IEEE (2019)
5. Awate, S.P., Garg, S., Jena, R.: Estimating uncertainty in MRF-based image segmentation: a perfect-MCMC approach. Med. Image Anal. **55**, 181–196 (2019)
6. Bert, A., et al.: An automatic method for colon segmentation in CT colonography. Comput. Med. Imaging Graph. **33**(4), 325–331 (2009)
7. Chen, D., Fahmi, R., Farag, A.A., Falk, R.L., Dryden, G.W.: Accurate and fast 3D colon segmentation in CT colonography. In: ISBI (2009)
8. Felzenszwalb, P.F., Huttenlocher, D.P.: Efficient graph-based image segmentation. Int. J. Comput. Vision **59**, 167–181 (2004)
9. Gayathri Devi, K., Radhakrishnan, R., et al.: Automatic segmentation of colon in 3D CT images and removal of opacified fluid using cascade feed forward neural network. Comput. Math. Methods Med. **2015** (2015)
10. Guachi, L., Guachi, R., Bini, F., Marinozzi, F., et al.: Automatic colorectal segmentation with convolutional neural network. Comput.-Aided Design Appl. **16**(5), 836–845 (2019)
11. Hanson, M.E., Pickhardt, P.J., Kim, D.H., Pfau, P.R.: Anatomic factors predictive of incomplete colonoscopy based on findings at CT colonography. Am. J. Roentgenol. **189**(4), 774–779 (2007)
12. Landman, B., Xu, Z., Igelsias, J., Styner, M., Langerak, T., Klein, A.: MICCAI multi-atlas labeling beyond the cranial vault–workshop and challenge. In: Proceedings of the MICCAI Multi-Atlas Labeling Beyond Cranial Vault-Workshop Challenge, vol. 5, p. 12 (2015)
13. Li, H., Xiong, P., An, J., Wang, L.: Pyramid attention network for semantic segmentation. arXiv preprint arXiv:1805.10180 (2018)
14. Lu, L., Zhang, D., Li, L., Zhao, J.: Fully automated colon segmentation for the computation of complete colon centerline in virtual colonoscopy. IEEE Trans. Biomed. Eng. **59**(4), 996–1004 (2011)
15. Malhotra, P., Gupta, S., Koundal, D., Zaguia, A., Enbeyle, W., et al.: Deep neural networks for medical image segmentation. J. Healthc. Eng. (2022)
16. Mohamad, M., Farag, A., Ali, A.M., Elshazly, S., Farag, A.A., Ghanoum, M.: Enhancing virtual colonoscopy with a new visualization measure. In: 2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018), pp. 294–297. IEEE (2018)
17. Nappi, J.J., Dachman, A.H., MacEneaney, P., Yoshida, H.: Effect of knowledge-guided colon segmentation in automated detection of polyps in CT colonography. In: Medical Imaging 2002: Physiology and Function from Multidimensional Images. SPIE (2002)
18. Oord, A.v.d., Li, Y., Vinyals, O.: Representation learning with contrastive predictive coding. arXiv preprint arXiv:1807.03748 (2018)
19. Ouyang, C., Biffi, C., Chen, C., Kart, T., Qiu, H., Rueckert, D.: Self-supervision with superpixels: training few-shot medical image segmentation without annotation. In: ECCV, Part XXIX 16, pp. 762–780. Springer, Cham (2020)
20. Rajamani, K., et al.: Segmentation of colon and removal of opacified fluid for virtual colonoscopy. Pattern Anal. Appl. **21**(1), 205–219 (2018)
21. Ramesh, K., Kumar, G.K., Swapna, K., Datta, D., Rajest, S.S.: A review of medical image segmentation algorithms. EAI Endors. Trans. Pervasive Health Technol. **7**(27), e6–e6 (2021)

22. Ravindran, Z., Das, N.S., et al.: Automatic segmentation of colon using multi-level morphology and thesholding. In: 2021 International Conference on Computer Communication and Informatics (ICCCI), pp. 1–4. IEEE (2021)
23. Ronneberger, O., Fischer, P., Brox, T.: U-net: convolutional networks for biomedical image segmentation. In: MICCAI 2015, Part III, pp. 234–241. Springer, Cham (2015)
24. Roy, A.G., Navab, N., Wachinger, C.: Recalibrating fully convolutional networks with spatial and channel squeeze and excitation blocks. IEEE Trans. Med. Imaging **38**(2), 540–549 (2018)
25. Roy, A.G., Siddiqui, S., Pölsterl, S., Navab, N., Wachinger, C.: squeeze & excite-guided few-shot segmentation of volumetric images. Med. Image Anal. **59**, 101587 (2020)
26. Sarkar, A., Biswas, M.K., Kartikeyan, B., Kumar, V., Majumder, K.L., Pal, D.: A MRF model-based segmentation approach to classification for multispectral imagery. IEEE Trans. Geosci. Remote Sens. **40**(5), 1102–1113 (2002)
27. Shamir, R.R., Duchin, Y., Kim, J., Sapiro, G., Harel, N.: Continuous dice coefficient: a method for evaluating probabilistic segmentations. arXiv (2019)
28. Tang, H., Liu, X., Sun, S., Yan, X., Xie, X.: Recurrent mask refinement for few-shot medical image segmentation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 3918–3928 (2021)
29. Wang, K., Liew, J.H., Zou, Y., Zhou, D., Feng, J.: PANet: few-shot image semantic segmentation with prototype alignment. In: proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 9197–9206 (2019)
30. Wasserthal, J., et al.: TotalSegmentator: robust segmentation of 104 anatomic structures in CT images. Radiol.: AI (2023)
31. Wu, H., Xiao, F., Liang, C.: Dual contrastive learning with anatomical auxiliary supervision for few-shot medical image segmentation. In: ECCV 2022, pp. 417–434. Springer, Cham (2022)
32. Zhang, C., Lin, G., Liu, F., Yao, R., Shen, C.: CANet: class-agnostic segmentation networks with iterative refinement and attentive few-shot learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 5217–5226 (2019)
33. Zhang, H., et al.: ResNest: split-attention networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 2736–2746 (2022)
34. Zhang, X., Wei, Y., Yang, Y., Huang, T.S.: SG-one: similarity guidance network for one-shot semantic segmentation. IEEE Trans. Cybern. **50** (2020)

Differential Diagnosis of Thyroid Tumors Through Information Fusion from Multiphoton Microscopy Images Using Fusion Autoencoder

Harshith Reddy Kethireddy[1(✉)], A. Tejaswee[1], Lucian G. Eftimie[2], Radu Hristu[3], George A. Stanciu[3], and Angshuman Paul[1]

[1] Indian Institute of Technology Jodhpur, Jodhpur, India
{reddy.16,tejaswee.1,apaul}@iitj.ac.in
[2] Central University Emergency Military Hospital, Bucharest, Romania
[3] Center of Microscopy-Microanalysis and Information Processing, National University of Science and Technology Politehnica Bucharest, Bucharest, Romania
radu.hristu@upb.ro, stanciu@physics.pub.ro

Abstract. Thyroid carcinomas are often diagnosed by histopathology, which is widely regarded as the most reliable method. However, alternative imaging modalities may also provide meaningful information about thyroid tumors. Multiphoton Microscopy (MPM) images may be one of them. MPM images include Second Harmonic Generation (SHG) and Two-Photon Excitation Fluorescence (TPEF) images. Nevertheless, the field of automated analysis of MPM images for the diagnosis of cancer is in its infancy. We propose a strategy for the differential diagnosis of thyroid tumors through information fusion from different types of MPM images. We introduce a novel fusion autoencoder (FAE) for this task. The fused information from the FAE is subsequently used by a classifier module for the differential diagnosis of thyroid tumors. Our method is one of the first approaches to look into the possibility of using MPM images for the diagnosis of thyroid tumors. Extensive experiments demonstrate the superiority of the proposed method compared to a number of state-of-the-art classification techniques. The code for the paper can be found at https://github.com/HarshithK13/ICPR2024-Thyroid-Diagnosis.git.

Keywords: Multiphoton Microscopy Images · Information Fusion · Fusion Autoencoder · Thyroid Tumor

1 Introduction

Thyroid tumors can be either benign (e.g., Follicular Adenoma (FA)) or malignant (e.g., Follicular Thyroid Carcinoma (FTC)). Histopathological analysis of

H. R. Kethireddy and A. Tejaswee—These authors contributed equally to this work.
This work is partially supported by SEED grant from IIT Jodhpur.

A. Antonacopoulos et al. (Eds.): ICPR 2024, LNCS 15313, pp. 80–93, 2025.
https://doi.org/10.1007/978-3-031-78201-5_6

the nodule is generally required for the identification of malignancies. Multiphoton Microscopy (MPM) imaging is an advanced imaging technique that utilizes nonlinear optical processes, such as Second Harmonic Generation (SHG) and Two-Photon Excited Fluorescence (TPEF), to provide high-resolution, three-dimensional images of biological tissues. SHG microscopy is useful for imaging collagen or myosin. TPEF is emitted by proteins in tissue which are autofluorescent. Depending on excitation and detection wavelengths, different tissue components can be imaged [1,2]. Thus, MPM can provide information about the tissues that may not be obtained using conventional histopathology. Backward-detected SHG (BSHG) imaging reveals a punctate pattern stemming from the presence of small-diameter, segmental collagen, facilitating the detection of fibrillogenesis in immature tissue. In contrast, both Forward-detected SHG (FSHG) and BSHG images of mature collagen fibrils display identical features. Thus, SHG microscopy specifically offers detailed visualization of collagen architecture inside a neoplasm [3], while TPEF imaging offers essential insights into intranodular details [4], making MPM a valuable tool in biomedical research, particularly for investigating tissue morphology and pathology. Thus, a combination of images from BSHG, FSHG, and TPEF images can provide complementary information about the sample under investigation.

Although SHG images are used for medical imaging applications [5], their use for the diagnosis of thyroid tumors is less explored. Currently, the thickness of the nodule capsule is the only significant histopathological characteristic that is associated with malignancy [6], as compared to benign nodules. It is crucial to distinguish thyroid carcinomas from adenomas and nodular goiters [7] since a misdiagnosis of this condition can have severe consequences, despite its low incidence rate. Considering corneal edema detection as mentioned in [5], SHG microscopy data has proven to be useful to identify seamlessly and monitor the architectural changes in the collagen of the cornea. It uses deep learning techniques to classify edematous corneal tissues using a combination of multiple models, and the model of such a combination has given better estimates than using stand-alone models, viz., ResNet-50 [8], InceptionV3 [9], and the Flexible Lightweight Model for Bioimage Analysis (FLIMBA). In [10], the authors show the benefits of using SHG images of ovarian tissues which get characterized using deep learning methods.

We propose a method for the differential diagnosis of thyroid tumors using MPM images of multiple modalities. We design a fusion autoencoder (FAE) that takes a stack of BSHG, FSHG and TPEF images for a Region of Interest (RoI) as inputs and provides feature maps with the fused information. Although many researchers have designed automated methods for the analysis of histopathology images from thyroid nodules [11], the use of MPM images in this context is relatively rare. Our primary contributions to this work are as follows:

- We design a method for the differential diagnosis of thyroid tumors using MPM images of three modalities.
- We propose a fusion autoencoder that can fuse information from BSHG, FSHG and TPEF images. The fused information is used for the classification of thyroid tumors.

– We experimentally demonstrate the utility of using MPM images from three different modalities.

The rest of the paper is organized as follows. We discuss the proposed method in Sect. 2 followed by experiments and results in Sect. 3. Finally, the paper is concluded in Sect. 5.

2 Methodology

We introduce a novel approach for the differential diagnosis of thyroid tumors based on MPM images, including BSHG, FSHG, and TPEF modalities. Our goal is to classify tumors into FA and FTC categories by fusing information from these images. These diverse modalities offer complementary insights into thyroid nodules [12]. Our objective is to integrate information extracted from BSHG, FSHG, and TPEF images of thyroid nodule capsules to classify tumors into FA and FTC categories. To achieve this, we propose a methodology comprising two main components: an Information Fusion Module (IFM) and a Classification Module (CM). The IFM, crucial for combining data from multiple modalities effectively, is devised around a novel fusion autoencoder architecture used for our specific diagnostic task. The fused information derived from the IFM feeds into the CM for the final diagnosis. A visual representation of our proposed method is illustrated in Fig. 1. This schematic encapsulates the integration of BSHG, FSHG, and TPEF images through our fusion autoencoder.

2.1 The Information Fusion Module (IFM)

In designing the IFM, several key considerations are taken into account. Firstly, it is of vital importance that the fusion process effectively fuses salient information from BSHG, FSHG, and TPEF images, each corresponding to an RoI. Autoencoders have demonstrated efficacy in extracting significant features from data [13]. Therefore, we propose employing a Fusion Autoencoder (FAE) for this task, aimed at extracting relevant information from the different modalities.

The FAE architecture, as depicted in Fig. 1, comprises of one input head and three output heads. Each RoI generates a stack of corresponding BSHG, FSHG, and TPEF images, which are then stacked as input to the FAE. The encoder within the FAE creates a latent space representation from the stacked images across these three modalities. Subsequently, the decoder consists of common layers, along with private layers dedicated to each of the three output heads. These private layers facilitate the preservation and propagation of modality-specific information. The private layers help to propagate modality-specific information. One output head is tasked to reconstruct the BSHG image solely from the latent space representation, while the other two output heads aim to reconstruct the FSHG and the TPEF images, respectively. Because of this, the fused feature maps derived from the latent space are expected to encapsulate a fusion of the salient information from all three imaging modalities.

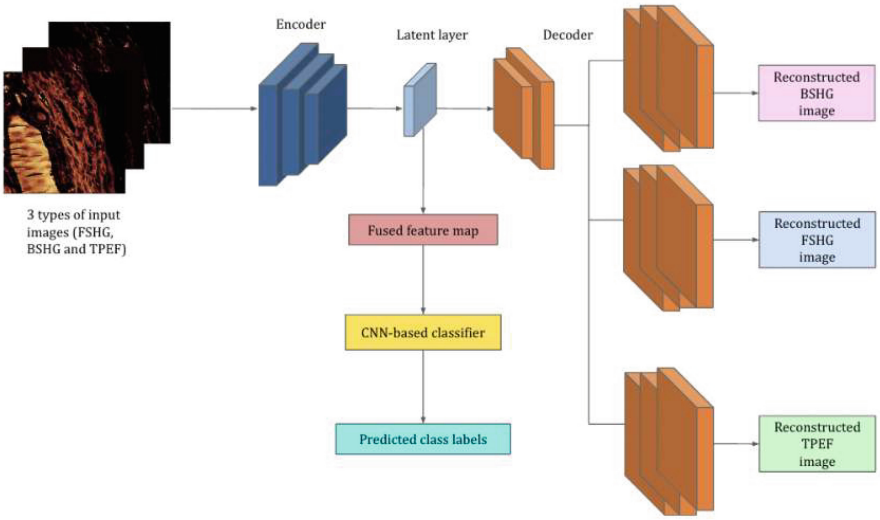


Fig. 1. A block diagram of the proposed method consisting of IFM (designed using FAE) and CM (designed using CNN-based classifier). A stack of BSHG, FSHG, and TPEF images of a single channel resulting in an image of 512 × 512 × 3-dimensional input is applied to the input layer of the encoder. The latent space representations are passed onto the decoder component that reconstructs the BSHG, FSHG, and TPEF images through three separate branches, respectively. L_B, L_F, L_T are losses from the reconstructed BSHG, FSHG, and TPEF images, respectively. The CNN-based classifier predicts the class label using the latent space representation (fused feature maps).

Let $I_B(n)$, $I_F(n)$, $I_T(n)$ be the BSHG, FSHG and TPEF images, respectively corresponding to RoI n. We create a stack

$$I(n) = \{I_B(n), I_F(n), I_T(n)\} \tag{1}$$

and apply $I(n)$ as input to the FAE. Let $I'_B(n)$, $I'_F(n)$ and $I'_T(n)$ represent the reconstructed BSHG, FSHG and TPEF images, respectively. We define the reconstruction loss for BSHG images (L_B) as the mean squared error between input BSHG images $I_B(\cdot)$ and the reconstructed BSHG images $I'_B(\cdot)$. Similar losses are defined for the FSHG images (L_F) and TPEF images (L_T). Therefore, the total loss for the FAE is

$$L = L_B + \alpha L_F + \beta L_T, \tag{2}$$

where α and β are the weightage of loss of the FSHG images and TPEF images relative to BSHG images. The weights for the different reconstruction terms are the hyperparameters of our model. These terms are chosen based on the validation performance. Our FAE is trained by minimizing L. The latent layer representations from the trained FAE is fed to the classification module.

2.2 The Classification Module (CM)

The CM is composed of a convolutional neural network-based classifier [14]. Using the latent space representations of the RoIs obtained from the IFM, the classifier undergoes training. This training process involves minimizing binary cross-entropy loss [15], which is designed for the task of two-class (FA and FTC) classification.

2.3 Inference

During the inference phase, a stack of BSHG, FSHG, and TPEF images corresponding to a test RoI is applied into the trained FAE. The FAE then generates a latent space representation from these images. This latent space representation is subsequently passed to the CM, where it is utilized to determine the final class labels for the given RoI.

2.4 Implementation Details

In our experimental setup, we conduct hyperparameter tuning for the FAE on a validation dataset comprising approximately 10% of the total data. The validation set is used to evaluate the performance of each combination. Through this process, we explored a range of hyperparameters and identified the set that achieved the highest mean AUROC score.

The proposed FAE has 25 convolutional layers, each followed by a GeLU activation function. Out of these, 12 are encoding layers, a latent layer and the other 12 are part of the decoder. The proposed architecture has 3×3 kernels in every convolutional layer. The decoder contains some common layers after the latent layer and gets branched out after a specific decoding layer (after the 6^{th} decoding layer in the experimental setup) to retain information about individual modalities as well as to have the fused information propagated further through these images. We use Max Pool layers of 2×2 dimensions after the first, and the sixth layers. In addition to these layers, we have two batch norm layers, one in the encoder and the other in the decoder components. The size of the feature maps after every convolutional layer starting from the first layer of encoder to the output of decoder are shown in Table 1.

The CM utilizes an EfficientNet-B4 architecture pretrained on ImageNet [16]. The FAE is trained for 30 epochs, while the classifier undergoes training for 50 epochs. We set the batch size for FAE input as 1 and 32 for the classifier. Both the FAE and the CNN-based classifier employ the Adam optimizer [17] with a learning rate of 0.0001. We use sigmoid activation function at the last classification layer. For a stack of input RoIs, we get class probabilities for the FA and FTC classes. The class with the highest probability score is considered to be the class of the input RoIs. For the ablation study, we maintain the same hyperparameter configurations, but find that utilizing LeakyReLU activation function yields better results for the fusion of BSHG with FSHG. Consequently, we adjust the learning rate for both the IFM and CM to 0.001. We tune the

Table 1. Feature Map Sizes After Each Convolutional Layer

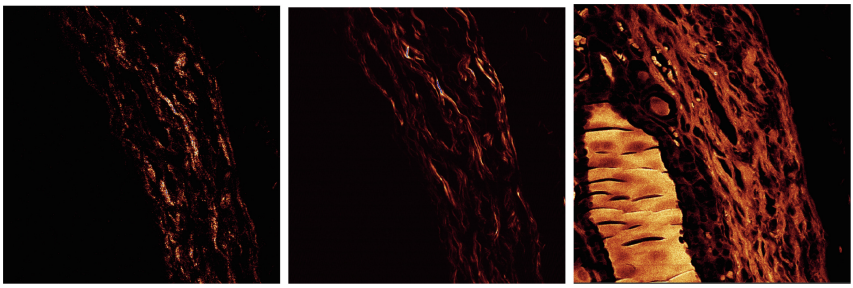
Layer	Batch Size	Output Channels	Height	Width
1	1	12	502	502
2	1	14	500	500
3	1	16	248	248
4	1	20	246	246
5	1	20	244	244
6	1	32	242	242
7	1	64	240	240
8	1	128	118	118
9	1	128	116	116
10	1	256	114	114
11	1	512	112	112
12	1	512	110	110
13	1	3	108	108
14	1	256	112	112
15	1	128	114	114
16	1	64	116	116
17	1	32	118	118
18	1	24	120	120
19	1	12	122	122
20	1	1	124	124
21	1	1	126	126
22	1	1	128	128
23	1	1	257	257
24	1	1	512	512
25	1	1	512	512

hyperparameters for individual classifiers using the validation data, identifying an optimal learning rate of 0.001 and a batch size of 16. We take $\alpha = 0.5$ and $\beta = 0.05$ in the loss function of (2).

3 Experiments and Results

3.1 Dataset

Our dataset contains different RoIs for 28 distinguishable tissue sections. We have 115 RoIs on a total of 8 tissue sections and 181 RoIs on 20 other tissue sections in FA and FTC categories, respectively. Further, each BSHG and FSHG images folder has raw data with ten linear polarization images captured at 0°



(a) (b) (c)

Fig. 2. Sample (a) BSHG, (b) FSHG and (c) TPEF images of linear polarization at 0° of a particular RoI

to 180° with an interval of 20°. Sample images of BSHG, FSHG and TPEF of a particular RoI are shown in Fig. 2. Tissue samples were obtained after partial or total thyroidectomy and were prepared according to standard histology protocols. Thin tissue sections stained with H&E were reviewed by a senior pathologist in order to place the diagnosis of either FA or FTC. Whole slide images were acquired from all the tissue sections. These virtual slides were annotated by the pathologist in order to highlight the nodule capsule surrounding the thyroid nodules, which were of interest in the present study. Using these annotated virtual slides as guidance, MPM images were collected on tissue slides around the thyroid nodule capsule. For each RoI, three images were simultaneously collected: BSHG, FSHG and TPEF each having dimensions of 512 × 512 pixels. No postprocessing was applied to the images acquired by the nonlinear optical microscope.

3.2 Comparative Performances

To ensure that the train, validation, and test sets don't have any overlapping RoIs from the same tissue section, we adopt a dataset splitting strategy based on tissue sections for our experiments. This approach prevents any information leakage across splits. We use 70:10:20 split of tissue sections for training, validation, and testing. A random split of the dataset may result in an unequal number of training RoIs from FA and FTC. To deal with that, we perform splitting such that the number of training RoIs from the two classes are almost the same. Also, for each RoI, out of 10 polarization images of each modality, we average the pixel values of 9 polarization images to streamline the data representation (excluding the 180° image as it is the same as the 0° image) to form a 512 × 512-dimensional resultant image.

We compare the proposed method with various state-of-the-art classifiers including ResNet-18, ResNet-50, DenseNet-121 [18], EfficientNet-B0 (ENet-B0), and EfficientNet-B4 (ENet-B4) using individual modality images and evaluate their performances. All of these classifiers are pretrained on the ImageNet

Table 2. Various performance metrics obtained using different classifiers on the test dataset for individual modalities (BSHG, FSHG and TPEF images) are captured in the table. They are compared with the metric values derived from the proposed architecture where all three modalities are stacked together as input.

Modality	Model	Class	Precision	Recall	F1 Score	AUROC
BSHG	ResNet-18	FA	0.53 ± 0.5	0.10 ± 0.1	0.16 ± 0.2	0.55 ± 0.1
		FTC	0.72 ± 0.0	0.99 ± 0.0	0.83 ± 0.0	
	ResNet-50	FA	0.60 ± 0.2	0.25 ± 0.2	0.30 ± 0.1	0.57 ± 0.1
		FTC	0.74 ± 0.0	0.89 ± 0.1	0.81 ± 0.0	
	DenseNet-121	FA	0.63 ± 0.2	0.43 ± 0.2	0.46 ± 0.1	0.63 ± 0.0
		FTC	0.78 ± 0.0	0.84 ± 0.1	0.80 ± 0.1	
	EfficientNet-B0	FA	0.45 ± 0.1	0.47 ± 0.2	0.40 ± 0.1	0.58 ± 0.1
		FTC	0.76 ± 0.0	0.67 ± 0.2	0.70 ± 0.1	
	EfficientNet-B4	FA	0.42 ± 0.3	0.37 ± 0.4	0.23 ± 0.2	0.51 ± 0.0
		FTC	0.57 ± 0.3	0.66 ± 0.4	0.59 ± 0.3	
FSHG	ResNet-18	FA	0.41 ± 0.3	0.38 ± 0.4	0.26 ± 0.2	0.56 ± 0.1
		FTC	0.77 ± 0.1	0.73 ± 0.3	0.69 ± 0.2	
	ResNet-50	FA	0.71 ± 0.3	0.18 ± 0.1	0.26 ± 0.1	0.56 ± 0.0
		FTC	0.73 ± 0.0	0.94 ± 0.1	0.82 ± 0.0	
	DenseNet-121	FA	0.54 ± 0.2	0.56 ± 0.2	0.49 ± 0.1	0.63 ± 0.1
		FTC	0.80 ± 0.0	0.69 ± 0.3	0.70 ± 0.2	
	EfficientNet-B0	FA	0.59 ± 0.2	0.52 ± 0.2	0.47 ± 0.1	0.63 ± 0.1
		FTC	0.79 ± 0.0	0.75 ± 0.2	0.75 ± 0.1	
	EfficientNet-B4	FA	0.63 ± 0.2	0.40 ± 0.1	0.45 ± 0.1	0.63 ± 0.0
		FTC	0.77 ± 0.0	0.85 ± 0.1	0.80 ± 0.1	
TPEF	ResNet-18	FA	0.36 ± 0.0	0.38 ± 0.0	0.37 ± 0.0	0.54 ± 0.0
		FTC	0.72 ± 0.0	0.70 ± 0.1	0.71 ± 0.0	
	ResNet-50	FA	0.34 ± 0.1	0.39 ± 0.1	0.35 ± 0.1	0.53 ± 0.0
		FTC	0.72 ± 0.0	0.69 ± 0.1	0.70 ± 0.1	
	DenseNet-121	FA	0.63 ± 0.2	0.43 ± 0.2	0.46 ± 0.1	0.63 ± 0.0
		FTC	0.78 ± 0.0	0.84 ± 0.1	0.80 ± 0.1	
	EfficientNet-B0	FA	0.40 ± 0.0	0.63 ± 0.2	0.47 ± 0.1	0.61 ± 0.1
		FTC	0.83 ± 0.1	0.58 ± 0.2	0.64 ± 0.2	
	EfficientNet-B4	FA	0.39 ± 0.0	0.72 ± 0.3	0.50 ± 0.1	0.63 ± 0.1
		FTC	0.85 ± 0.1	0.54 ± 0.1	0.65 ± 0.1	
Proposed	EfficientNet-B4	FA	0.72 ± 0.2	0.47 ± 0.3	0.50 ± 0.2	0.66 ± 0.1
		FTC	0.74 ± 0.1	0.85 ± 0.1	0.78 ± 0.1	

Table 3. AUROC (mean ± sd) over ten runs for the proposed method. This table also shows the results using images of individual modalities and images from different combinations of two modalities.

Modality	Best Model	AUROC
BSHG	DenseNet-121	0.57 ± 0.07
FSHG	EfficientNet-B4	0.56 ± 0.07
TPEF	DenseNet-121	0.58 ± 0.07
BSHG-FSHG	EfficientNet-B4	0.53 ± 0.04
BSHG-TPEF	EfficientNet-B4	0.53 ± 0.05
FSHG-TPEF	EfficientNet-B4	0.53 ± 0.04
Proposed	EfficientNet-B4	0.66 ± 0.10

dataset. For each test data point, we compute the probabilities of belonging to classes FA and FTC. The data point is then assigned the class label with the higher probability. Based on these predicted class labels and the ground truth class labels, the values of the recall and precision are calculated. We calculate a recall and a precision value considering FA as the positive class. We do the same considering FTC as the positive class. This enables us to get the class-wise recall and precision values. We run each method for ten times. For each run, the learnable parameters are initialized randomly. We take the best five runs out of ten runs to rule out the possibility of very poor initialization. The results for best five runs (mean ± sd) are presented in Table 2. Notice that our method outperforms all competitors in terms of the mean AUROC. Subsequently, we look into the performances of competing methods when presented with stacked images from three modalities as input to the modified autoencoder. This is a type of early fusion [19]. This fusion helps to combine the information from the distinct sets of images and hence improves the capabilities of the model, thus leading to a more robust and efficient solution. Table 4 displays the metric values when computed on the images using early fusion technique and passing the latent layer feature maps to every competing classifier for comparison. It can be observed that our proposed method with ENet-B4 shows superior performances compared to its competitors. Results on sample images using the proposed method are presented in Fig. 3.

Additionally, Table 3 shows the AUROC scores (mean ± sd) of the proposed method over ten runs. This table also contains the results over ten runs using images of individual modalities and a combination of images from two modalities. For these experiments, we take the classifiers that provided the best results when top five runs are considered.

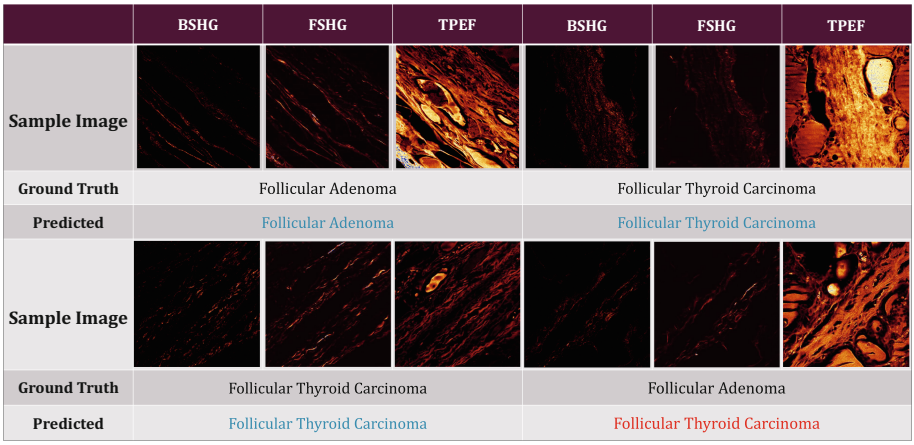


Fig. 3. Sample BSHG, FSHG and TPEF images with their ground truth and predicted class labels using our method (blue: correct prediction, red: incorrect prediction). (Color figure online)

3.3 Ablation Studies

We perform ablation studies to examine the impact of information fusion using images from three modalities. To this end, we perform experiments with different combinations of modalities to construct stacked images at the input of our method. Figure 4 illustrates the results obtained with various combinations of two modalities, namely, BSHG-FSHG, FSHG-TPEF, and BSHG-TPEF, alongside the results obtained using our proposed method. When utilizing two modalities, we incorporate two output heads for the ablation studies. It is evident from the results that using images from any two modalities leads to inferior performance compared to our proposed method. This shows the importance of information fusion using images from all three modalities in achieving optimal classification performance.

4 Discussion

As mentioned before, both BSHG and FSHG images primarily provide information on collagen or myosin. On the other hand, since TPEF is emitted by proteins in tissues which are autofluorescent, TPEF may provide more complementary information when combined with either BSHG or FSHG images.

Hence, our results shown in Table 3 indicate that the combination of BSHG and TPEF or FSHG and TPEF as input leads to better performance compared to using BSHG and FSHG alone. Specifically, the proposed method that integrates all three modalities-BSHG, FSHG, and TPEF-achieves the highest performance with a maximum mean AUROC score of 0.66.

Table 4. Performance metrics with different classifiers that take the latent space representation of our autoencoder as input.

Model	Class	Precision	Recall	F1 Score	AUROC
ResNet-18	FA	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.5 ± 0.00
	FTC	0.63 ± 0.00	1.00 ± 0.00	0.77 ± 0.00	
ResNet-50	FA	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.5 ± 0.00
	FTC	0.63 ± 0.00	1.00 ± 0.00	0.77 ± 0.00	
DenseNet-121	FA	0.27 ± 0.26	0.30 ± 0.38	0.24 ± 0.27	0.53 ± 0.06
	FTC	0.66 ± 0.05	0.77 ± 0.35	0.66 ± 0.22	
EfficientNet-B0	FA	0.28 ± 0.26	0.39 ± 0.38	0.33 ± 0.31	0.57 ± 0.08
	FTC	0.71 ± 0.09	0.75 ± 0.23	0.7 ± 0.07	
Proposed	FA	0.72 ± 0.19	0.47 ± 0.27	0.50 ± 0.22	0.66 ± 0.10
	FTC	0.74 ± 0.09	0.85 ± 0.12	0.78 ± 0.05	

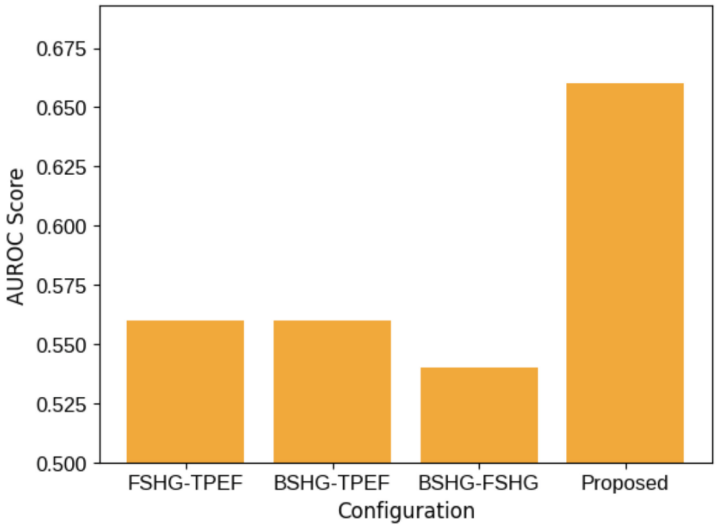


Fig. 4. Mean AUROC scores over five best runs using different combinations of modalities (FSHG with TPEF, BSHG with TPEF and BSHG with FSHG) alongside the proposed method (BSHG, FSHG, and TPEF as input).

Furthermore, the proposed method outperforms models that use any two modalities in combination. For example, the combination of BSHG and FSHG achieved a lower mean AUROC score compared to when TPEF was included. This combination hence enhances the capability of the model to distinguish between different classes of thyroid tumors.

Our ablation studies shown in Fig. 4 also show the significance of using three modalities. This finding highlights the potential of multimodal approaches in

medical imaging, where different imaging techniques can complement each other to provide a more comprehensive understanding of the tissue characteristics. The synergy between the network architecture and the new data modalities is a key factor driving the observed performance gains.

5 Conclusion

We delve into the potential of information fusion from different types of MPM images for the differential diagnosis of thyroid tumors. To achieve this, we design a Fusion Autoencoder, aimed at integrating information from three distinct modalities. The latent space representation of the autoencoder is found to provide meaningful information through the fusion of MPM images of three different modalities. Rigorous experiments show that the proposed method can obtain a mean AUROC score of 0.66. However, the use of individual MPM images can achieve a maximum mean AUROC score of 0.63. This shows the impact of information fusion in our model. Furthermore, ablation studies show that the use of information fusion from images of any two modalities is significantly less effective compared to the proposed strategy of using images from three modalities. The present work is a proof-of-concept study to look into the utility of multiphoton microscopy images. This method of fusing information from images of multiple modalities and the use of a Fusion Autoencoder shows promising results for the diagnosis of thyroid tumors. In the future, we will explore the feasibility of integrating information from histopathological images with the information from MPM images to further enhance the accuracy of thyroid tumor diagnosis. We will also look into the possibility of utilizing larger datasets for our experiments. Moreover, we intend to extend the application of such information fusion strategies to incorporate non-image medical data, thereby broadening the scope of diagnostic capabilities.

References

1. Li, L.Z., et al.: Two-photon autofluorescence imaging of fixed tissues: feasibility and potential values for biomedical applications. In: Oxygen Transport to Tissue XLI, pp. 375–381. Springer (2020)
2. Jun, Y.W., et al.: Addressing the autofluorescence issue in deep tissue imaging by two-photon microscopy: the significance of far-red emitting dyes. Chem. Sci. **8**(11), 7696–7704 (2017)
3. Campagnola, P.J., Dong, C.Y.: Second harmonic generation microscopy: principles and applications to disease diagnosis. Laser Photonics Rev. **5**, 13–26 (2011). https://doi.org/10.1002/lpor.200910024
4. Mulligan, S.J., Garrod, B.D., Leake, M.A.: Two-photon fluorescence microscopy: basic principles, advantages, and risks. Methods Cell Biol. **86**, 105–129 (2007). https://doi.org/10.1016/S0091-679X(06)86010-3
5. Anton, S.R., et al.: Automated detection of corneal edema with deep learning-assisted second harmonic generation microscopy. IEEE J. Sel. Top. Quantum Electron. **29**(6: Photonic Signal Processing), 1–10 (2023). https://doi.org/10.1109/JSTQE.2023.3149295

6. Volante, M., Papotti, M.: A practical diagnostic approach to solid/trabecular nodules in the thyroid. Endocrine Pathol. **19**, 75–81 (2008)
7. Cooper, D.S., et al.: Revised American Thyroid Association management guidelines for patients with thyroid nodules and differentiated thyroid cancer: the American Thyroid Association (ATA) guidelines taskforce on thyroid nodules and differentiated thyroid cancer. Thyroid **19**(11), 1167–1214 (2009)
8. Sarwinda, D., et al.: Deep learning in image classification using residual network (ResNet) variants for detection of colorectal cancer. Procedia Comput. Sci. **179**, 423–431 (2021). https://doi.org/10.1016/j.procs.2021.01.025
9. Szegedy, C., et al.: Rethinking the inception architecture for computer vision. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2818–2826. IEEE (2016)
10. Wang, G., et al.: Automated ovarian cancer identification using end-to-end deep learning and second harmonic generation imaging. IEEE J. Sel. Top. Quantum Electron. **29**(4: Biophotonics), 1–9 (2023). https://doi.org/10.1109/JSTQE.2022.3228567
11. Eftimie, L.G., et al.: Differential diagnosis of thyroid nodule capsules using random forest guided selection of image features. Sci. Rep. **12**, 25788 (2022). https://doi.org/10.1038/s41598-022-25788-w
12. Hristu, R., et al.: PSHG-TISS: a collection of polarization-resolved second harmonic generation microscopy images of fixed tissues. Scientific Data **9**(1), 376 (2022). https://doi.org/10.1038/s41597-022-01201-3
13. Petscharnig, S., Lux, M., Chatzichristofis, S.: Dimensionality reduction for image features using deep learning and autoencoders. In: 2017 IEEE International Conference on Systems, Man and Cybernetics, pp. 1–6 (2017). https://doi.org/10.1145/3095713.3095737
14. Sarvamangala, D.R., Kulkarni, R.V.: Convolutional neural networks in medical image understanding: a survey. Evol. Intell. **15**(1), 1–22 (2022). https://doi.org/10.1007/s12065-021-00426-4
15. Ruby, U., Yendapalli, V.: Binary cross entropy with deep learning technique for image classification. Int. J. Adv. Trends Comput. Sci. Eng. **9**(10) (2020)
16. Deng, J., et al.: Imagetet: a large-scale hierarchical image database. In: 2009 IEEE conference on computer vision and pattern recognition, pp. 248–255. IEEE (2009)
17. Bock, S., Weiß, M.: A proof of local convergence for the Adam optimizer. In: 2019 international joint conference on neural networks (IJCNN), pp. 1–8. IEEE (2019)
18. Huang, G., Liu, Z., Weinberger, K.Q.: Densely connected convolutional networks. CoRR, abs/1608.06993 (2016)
19. Pereira, L.M., Salazar, A., Vergara, L.: A comparative analysis of early and late fusion for the multimodal two-class problem. IEEE Access **11**, 84283–84300 (2023). https://doi.org/10.1109/ACCESS.2023.3296098
20. Chen, X., Nadiarynkh, O., Plotnikov, S., Campagnola, P.J.: Second harmonic generation microscopy for quantitative analysis of collagen fibrillar structure. Nat. Protocols **7**(4), 654–669 (2012)
21. Tilbury, K., Hocker, J., Wen, B.L., Sandbo, N., Singh, V., Campagnola, P.J.: Second harmonic generation microscopy analysis of extracellular matrix changes in human idiopathic pulmonary fibrosis. J. Biomed. Opt. **19**(8), 086014 (2014)
22. Adur, J., et al.: Second harmonic generation microscopy as a powerful diagnostic imaging modality for human ovarian cancer. J. Biophotonics **7**(1–2), 37–48 (2014)
23. Pak, A., et al.: Comparative analysis of deep learning methods of detection of diabetic retinopathy. Cogent Eng. **7**(1), 1805144 (2020). https://doi.org/10.1080/23311916.2020.1805144

24. Tan, M., Le, Q.V.: EfficientNet: rethinking model scaling for convolutional neural networks. CoRR, abs/1905.11946 (2019). http://arxiv.org/abs/1905.11946
25. Pena, A.M., et al.: In vivo multiphoton multiparametric 3D quantification of human skin aging on forearm and face. Sci. Rep. **12**(1), 14863 (2022). https://doi.org/10.1038/s41598-022-07325-5
26. Stanciu, S.G., et al.: Toward next-generation endoscopes integrating biomimetic video systems, nonlinear optical microscopy, and deep learning. Biophys. Rev. **4**(2) (2023). https://doi.org/10.1007/s41048-023-00222-1
27. Gunes, H., Piccardi, M.: Affect recognition from face and body: early fusion vs. late fusion. In: 2005 IEEE International Conference on Systems, Man and Cybernetics, vol. 4, pp. 3437–3443 (2005). https://doi.org/10.1109/ICSMC.2005.1571679

Investigating the ABCDE Rule in Convolutional Neural Networks

Federico Bolelli(✉), Luca Lumetti, Kevin Marchesini, Ettore Candeloro, and Costantino Grana

Università degli Studi di Modena e Reggio Emilia, Modena, Italy
{federico.bolelli,luca.lumetti,kevin.marchesini,
ettore.candeloro,costantino.grana}@unimore.it

Abstract. Convolutional Neural Networks (CNNs) have been broadly employed in dermoscopic image analysis, mainly due to the large amount of data gathered by the International Skin Imaging Collaboration (ISIC). But where do neural networks look? Several authors have claimed that the ISIC dataset is affected by *strong biases*, *i.e.,* spurious correlations between samples that machine learning models unfairly exploit while discarding the useful patterns they are expected to learn. These strong claims have been supported by showing that deep learning models maintain excellent performance even when "no information about the lesion remains" in the debased input images. With this paper, we explore the interpretability of CNNs in dermoscopic image analysis by analyzing which characteristics are considered by autonomous classification algorithms. Starting from a standard setting, experiments presented in this paper gradually conceal well-known crucial dermoscopic features and thoroughly investigate how CNNs performance subsequently evolves. Experimental results carried out on two well-known CNNs, EfficientNet-B3, and ResNet-152, demonstrate that neural networks autonomously learn to extract features that are notoriously important for melanoma detection. Even when some of such features are removed, the others are still enough to achieve satisfactory classification performance. Obtained results demonstrate that literature claims on *biases* are not supported by carried-out experiments. Finally, to demonstrate the generalization capabilities of state-of-the-art CNN models for skin lesion classification, a large private dataset has been employed as an additional test set.

Keywords: ABCDE Rule · Convolutional Neural Networks · Skin Lesion Classification · Dataset Bias · Transfer Learning

1 Introduction

Skin cancer is the most common form of human cancer and a major public health issue. Malignant melanoma, although less common, is responsible for most of the deaths [6]. The early detection of skin cancer remains one of the key factors in preventing its progression to advanced stages and lowering mortality rates [39]. To do so, many dermatologists rely on dermoscopy, which is a

A. Antonacopoulos et al. (Eds.): ICPR 2024, LNCS 15313, pp. 94–111, 2025.
https://doi.org/10.1007/978-3-031-78201-5_7

form of in-vivo skin surface microscopy performed using special equipment to enhance the visibility of the pigmentation of the lesion and perform a faster, more accurate diagnosis over time. Unfortunately, dermoscopy image analysis must be performed by expert clinicians to be effective, and this is why many efforts have been made toward the creation of tools to assist non-specialized physicians in the analysis of dermoscopic images [2]. The outstanding results of deep learning in many different research areas [5,23,25,46], make it one of the most employed and effective options for analyzing medical images. However, the great discriminative power of neural networks comes at the cost of very low explainability. Hence, it is extremely difficult to understand the reasoning behind a model prediction [22,37], and this characteristic can also lead to the possibility of CNNs learning a *bias*. A bias can exist in different shapes and forms and may originate from different sources [34,45], but in the analysis carried out in this paper, we focus on *data-to-algorithm* biases, which, when used by machine-learning training algorithms, might result in biased algorithmic outcomes. In particular, a *dataset bias* can be defined as a collection of features that are semantically irrelevant to the investigated task, but which can be (undesirably) exploited by neural networks to improve the evaluation metrics, hindering their generalization capabilities [31]. This phenomenon has been thoroughly investigated by several authors [4,20] and our goal is to explore it in dermoscopic image analysis [19]. It is desirable for automatic skin lesion classification algorithms to focus on medically relevant features instead of considering irrelevant artifacts (*e.g.,* checkerboard patterns introduced by sharpening filters, black round borders, pen drawings, rulers, and hair) which should be ignored for classification.

The most common dermoscopic relevant features for melanoma detection, explicitly outlined by expert practitioners, are defined in the so-called "ABCDE rule": lesion *Asymmetry*, *Border* irregularity, *Color* variegation, *Diameter* (>6 mm), and *Evolution* over time [39].

Our study investigates how the performance of CNNs correlates with established dermoscopic criteria by methodically altering images to omit each of the ABCDE melanoma indicators. By selectively "removing" these elements, the research aims to discern the extent to which CNNs rely on authentic clinical features versus incidental image attributes. The contributions of this paper can be summarized as follows:

i) Making use of state-of-the-art interpretability tools, we examine the correlation between deep learning algorithms and well-known dermoscopic features (ABCDE rule) used by expert practitioners to perform diagnoses;
ii) We propose an extensive set of experiments to highlight how the discriminative power of state-of-the-art CNNs is affected by different dermoscopic features and verify the literature claims on dermoscopic datasets biases;
iii) We validate the generalization capabilities of state-of-the-art CNN algorithms for skin lesion classification by carrying out experiments on two totally distinct datasets: the combined ISIC2019 and ISIC2020 and a privately owned one that has no intersection with the former.

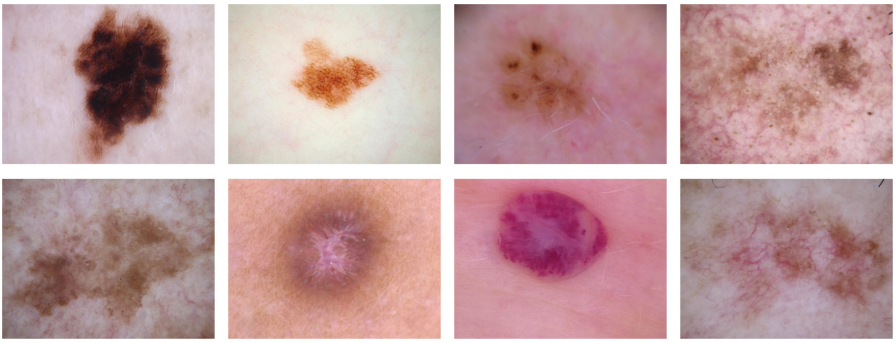


Fig. 1. Samples of the 2019 ISIC dataset. From left to right, top to bottom: Melanoma, Melanocytic Nevus, Basal Cell Carcinoma, Actinic Keratosis, Benign Keratosis, Dermatofibroma, Vascular Lesion, and Squamous Cell Carcinoma.

2 Related Work

CNNs have become the dominant machine learning approach, and the *scaling up* strategy [23] has been widely used to achieve accuracy results similar to those of dermatologists, particularly in skin lesion classification [1,16,17,26,32,33,38,51] and to aid in diagnosis even on low-resolution non-dermoscopic images [15]. However, despite their success, concerns about CNN focusing on irrelevant artifacts were highlighted by [30,40], where studies on multiple COVID-19 datasets, performed hiding sensitive information with large black squares, showed state-of-the-art networks focusing on dataset-specific features rather than clinically relevant ones, highlighting the incompatibility of those models for clinical usage.

In dermatology, Bissoto *et al.* [4] showed the effects of performing skin lesion classification while occluding the actual skin lesion with large black bounding boxes, obtaining a melanoma/non-melanoma classification AUC (Area Under the ROC Curve) score of 77.4%, which is quite inferior compared to state-of-the-art methods, but higher than what expert dermatologists can do [7], highlighting a potential reliance on non-diagnostic features. Additional studies confirmed the CNNs' learned filters focusing on both relevant features (*e.g.,* borders, and colors) and extraneous features (like artifacts surrounding the lesion) [3,53].

Autonomous systems in medical applications aim to act as support tools for clinicians and, therefore, must be trustworthy and highly interpretable. To aid in this task, the outcome explainability of neural networks can be increased thanks to several visualization strategies, like *CAM* (Class Activation Mapping), which have been proposed for the identification of image regions that most contribute to the final prediction [42,48].

In this paper, we make use of state-of-the-art interpretability tools, along with quantitative results, to examine the correlation between deep learning algorithms and well-known dermoscopic features [28] introduced in Sect. 1.

Table 1. Class distribution of the three employed datasets: 2019 and 2020 ISIC datasets and private dataset.

Class	Label	ISIC2019 %	ISIC2020 %	Private %
Melanoma	MEL	17.8	1.8	16.7
Melanocytic Nevus	NV	50.8	15.7	58.1
Basal Cell Carcinoma	BCC	13.0	–	7.6
Actinic Keratosis	AK	3.0	–	1.6
Benign Keratosis	BKL	10.0	0.7	6.3
Dermatofibroma	DF	0.9	–	1.0
Vascular Lesion	VASC	1.0	–	0.0
Squamous Cell Carcinoma	SCC	2.4	–	1.8
Unknown	unknown	–	81.9	6.9
Total		25 331	33 126	25 849

3 Dermoscopic Images

ISIC. The International Skin Imaging Collaboration (ISIC) began to aggregate a large-scale, publicly available collection of dermoscopic skin lesion images (Fig. 1) starting in 2016, with the aim of supporting research towards enhancing machine learning algorithms for automated skin cancer analysis, showcasing the results of researchers in several challenges and workshops hosted over the years [14]. The 2019 version of the ISIC archive contains a total amount of 25 331 labeled dermoscopy images, belonging to nine different classes [24], which represent eight types of skin lesion plus an additional category, not available in the training partition and containing dermoscopic images of different natures with respect to the other eight classes.

The available data is heavily imbalanced in classes, therefore the 2019 challenge official metric was the balanced accuracy, computed as the average sensitivity among classes regardless of their occurrence in the test set.

The successive 2020 SIIM-ISIC challenge dataset [41] gained patient-level contextual information, providing for each image an identifier that allows lesions from the same patient to be mapped to one another. This additional knowledge is frequently used by clinicians to diagnose melanoma and is especially useful in ruling out false positives in patients with many atypical nevi, leveraging the "ugly duckling sign" rule [18]. The challenge edition, hosted on Kaggle,[1] switched to a binary classification problem: benign or malignant, employing the AUC evaluation metric. In the subsequent sections of the paper, the name ISIC19-20 will be used to refer to the combination of ISIC2019 and ISIC2020 datasets. More details about such a combination are provided in Sect. 5. Table 1 summarizes ISIC dataset features.

Private Dataset. In order to evaluate the generalization capabilities of state-of-the-art CNNs models, we extend the experiments by means of a private der-

[1] kaggle.com/c/siim-isic-melanoma-classification.

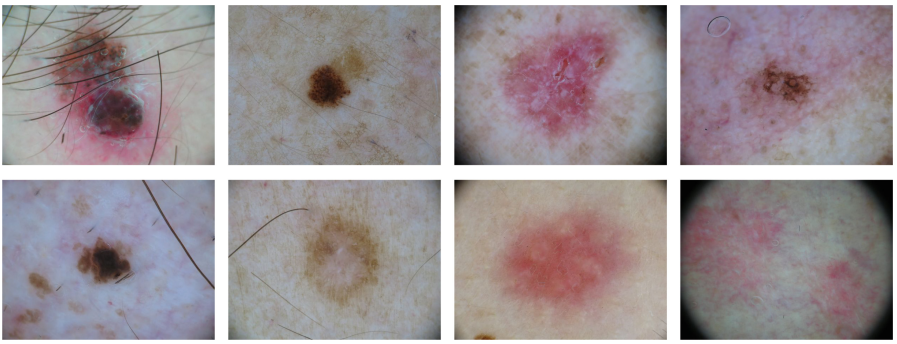


Fig. 2. Samples of the Private dataset. From left to right, top to bottom: Melanoma, Melanocytic Nevus, Basal Cell Carcinoma, Actinic Keratosis, Benign Keratosis, Dermatofibroma, Vascular Lesion, and Squamous Cell Carcinoma.

moscopic dataset (Fig. 2) consisting of 25 849 images, collected between 2003 and 2019 in the University Hospital of Modena using several distinct acquisition tools, and employing the same classes mapped into the ISIC2019 dataset.[2] This dataset presents a different category distribution compared to the ISIC2020 dataset, with a higher percentage of melanoma cases (Table 1). Similar to both ISIC datasets, the private collection of data contains several clinical information such as sex, age, and site of the lesion. Contrary to the public ISIC dataset, visual artifacts that could be considered a source of biases, such as rulers, ink markings/staining, and colored patches, are almost completely absent in our private dataset (7% ruler, 1.9% ink, and no images with patches). The whole set of dermoscopic images is used as an additional test set for the experiments and analyses carried out in this paper, and thus yields important information about the generalizability of state-of-the-art CNNs models and their possible application in real-world scenarios.

4 Investigating ABCDE Features

Neural networks for skin lesion classification have been shown to focus on relevant features for dermoscopic image analysis, aligned with the ABCDE rule [28,39], but they might also focus on irrelevant visual aspects that are common in malignant skin lesion images, such as artifacts related to pen drawings, markers, colored patches, or rulers. Moreover, additional research showed that CNNs are able to recognize acquisition device models and calibration settings, thus identifying the provenience of an image that might be highly related to the final diagnosis [30]. Hence, it is extremely important to be able to interpret which image characteristics neural networks take into account when making a class prediction to highlight potential *data-to-algorithm* biases. This can be achieved by means

[2] The dataset is currently under review by the ethical committee to be publicly released. After approval, it will be accessible at https://ditto.ing.unimore.it/.

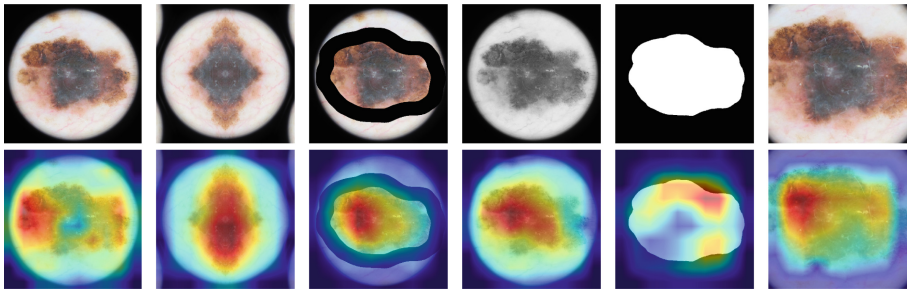


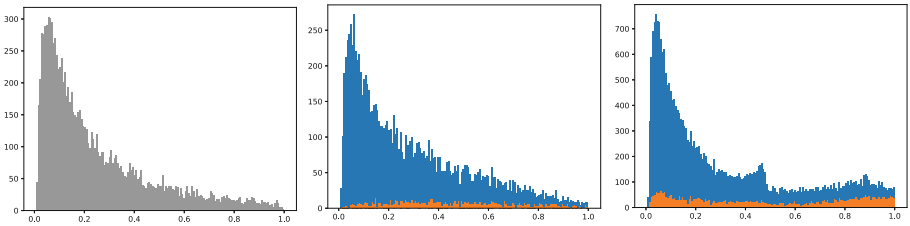
Fig. 3. Grad-CAM visualization when debasing different ABCD(E) properties. (a) Original, (b) Asymmetry, (c) Borders, (d) Grayscale, (e) Mask and (f) Diameter.

of Class Activation Mapping (CAM) strategies, employed in this paper. In particular, Grad-CAM [42] was exploited to locate the regions of an image that most contribute to the final prediction. We run an extensive set of experiments to study how introducing noise in the ABCDE properties affects neural network performance and analyze which sections of an image CNNs focus on when crucial features are debased or removed. Some of the experiments described in the following exploit segmentation masks obtained by means of DeepLabv3+ [12], trained using the 2017 ISIC segmentation task dataset [9]. Sample images obtained through feature debasing are reported in Fig. 3 and generated by means of the European Computer Vision Library (ECVL) [10]. Additional examples of ABCDE features debasing images can be found in Fig. 7 at the end of the paper.

By applying the feature debasing process described in the following of this section, we obtain five additional variations of each considered dataset (*i.e.,* five variations of the ISIC datasets and five variations of the Private dataset), each of them is employed for both training (ISIC19-20) and testing (ISIC19-20 and private dataset) selected models.

Tampering with Asymmetry. Asymmetry is one of the most important visual features for melanoma detection [35], it can be described as the difference in volume and shape of two parts of a skin lesion, obtained by *cutting* it with a straight line passing through its center. In order to train a symmetry-agnostic neural network, dermoscopic images can be split by a random straight line and by its perpendicular, both passing through the center of the lesion. Subsequently, a quarter of the image can be flipped over both axes to obtain a version of the original lesion with increased symmetry. Practically, the center of the image is aligned with the centroid of the lesion obtained from the segmentation mask. The image is then randomly rotated, and the top-right quarter is flipped with respect to the horizontal and vertical axes (Fig. 3b).

Concealing Borders. With the aim of removing valuable information about the shape of a skin lesion edge, which is a crucial aspect when assessing its



(a) (b) (c)

Fig. 4. Histograms of foreground density distribution within different test sets. (a) ISIC2020 official test set, (b) ISIC19-20 "Internal" test set, (c) Private Dataset. Benign and malignant skin lesions are depicted in blue and orange, respectively.

malignancy, we cover borders with a thick black line obtained from the contour of the segmentation map. Firstly, a morphological dilation operation is applied to the contour, with a kernel size proportional to both the image size and the foreground-background ratio. Then, to smooth out irregular segments, a Gaussian filter with a large kernel is applied. Finally, the black border image is superimposed on the original one, thus removing any information about the actual transition from the human skin (background) to the actual lesion (foreground). Figure 3c showcases an example of the end result image.

Removing Color. The presence of multiple colors within a single mole (blue, black, white, red, and brown) or the uneven distribution of color can sometimes be a warning sign of melanoma, since most benign lesions are usually a single shade of brown or tan [29]. Two different sets of experiments are conducted in order to assess the effects of discarding information about color from dermoscopic images. The first one is run by simply converting the image from RGB to grayscale, thus removing any knowledge about the different colors within a skin lesion (Fig. 3d). However, while this processing step erases any data about hue and saturation, it does not affect the luminance, thus leaving the CNN the chance to learn valuable features from the color distribution within moles. In the second experiment, color features are completely removed as we train a neural network to classify skin lesions using only their segmentation masks (Fig. 3e). In this extreme setup, the neural network is fed with minimal knowledge about skin lesions, and is forced to make a prediction based uniquely on noisy, automatically obtained, binary mole shapes.

Altering Diameter. Because skin cancer cells grow abnormally fast, diameter is one of the most important parameters in skin lesion classification. Unfortunately, dermoscopic images are acquired at several scales, which are not always included as metadata and can not be deducted from the image, as only a limited amount of samples contain a ruler. As a matter of fact, a mole that exceeds the borders of the image is not necessarily larger than one that does not; information about diameter is extremely noisy and very hard to investigate, yet poten-

tially extremely important for melanoma detection algorithms. To remove any information about mole dimensions, the foreground-background ratio of images is set to a fixed percentage. We choose to train a CNN uniquely with samples where the skin lesion represents the 80% of the image, as it yields good qualitative results. In order to achieve this, samples where the skin lesion is contained in just a small portion of the image must be cropped, using the mole centroid as the center, whereas images with moles covering more than 80% of the original sample are padded by reflecting the sections of the image closest to the borders. An example of the result of this process is illustrated in Fig. 3f. Additionally, the foreground density histograms in Fig. 4 show that this method mostly results in crops, whereas only a very small portion of the dataset (foreground percentage >80%) needs padding. By following the aforementioned *Crop&Pad* technique, in the rare cases of very elongated lesions, a small part of the mole is left out by the crop. However, each image will present the same number of foreground pixels, thus eliminating the image scale differences and bringing all the lesions to the same size.

About Evolving. Dermoscopic images seldom contain information about evolution, as only in a few cases follow-up data is provided in the existing datasets. Introducing such additional data in dermoscopic datasets would certainly have significant implications in the research field, but it cannot be considered and analyzed nowadays. For this reason, we were unable to experiment on lesion evolution.

5 Experiments

Datasets Preprocessing. To harmonize the two ISIC datasets, the following pre-processing steps have been employed: first, the 2020 classes are mapped into 2019 ones; then, in order to compensate for dissimilarities between image sizes, a squared center crop is performed to produce images of $min(h, w) \times min(h, w)$ pixels, later resized to 768×768. The combined ISIC2019 and ISIC2020 datasets, purged of duplicates [50], provide a total of 57 964 images. As previously mentioned, we refer to this combined dataset as ISIC19-20. As introduced in Sect. 3, we also employed a private dataset with the same class mapping as the ISIC19-20.

Networks and Training Details. Our study utilizes two of the networks constituting the ensemble strategy adopted by the ISIC2020 Kaggle challenge winner [21], *i.e.,* EfficientNet-B3 and ResNet-152. While achieving performances that are comparable with the state-of-the-art, they have a limited computational load in terms of time and memory and allow us to perform the extensive set of experiments described in this section. Input image sizes are 300×300 and 256×256 for the two models, respectively. Both networks are trained with the Cross-Entropy loss and Adam optimizer [27], with a learning rate of 3×10^{-5}.

Table 2. Experimental results obtained by training (and testing) the models on the input configurations described in Sect. 4. Each of the *Experiment* correspond to a training performed on the corresponding debased ISIC19-20 dataset and tested on the debased ISIC19-20 "internal" test set. Threshold is set to 0.5.

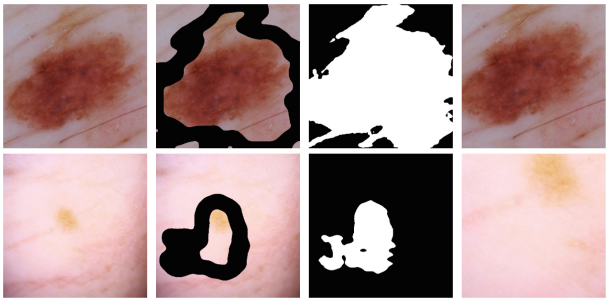
Model	Experiment	AUC ROC	Precision	Recall (Sensitivity)	Specificity	F1-Score	Accuracy
EfficientNet-B3	Original	0.9671	0.7821	0.7180	0.9808	0.7487	0.9577
	Asymmetry	0.9448	0.7755	0.5399	0.9850	0.6366	0.9459
	Borders	0.9605	0.7326	0.6678	0.9766	0.6987	0.9495
	Color (Grayscale)	0.9559	0.7420	0.7071	0.9763	0.7241	0.9527
	Color (Mask)	0.8017	0.6897	0.0656	0.9972	0.1198	0.9154
	Diameter	0.9724	0.8216	0.7399	0.9845	0.7786	0.9631
ResNet-152	Original	0.9572	0.7548	0.6934	0.9782	0.7228	0.9531
	Asymmetry	0.9188	0.6539	0.4848	0.9837	0.5568	0.9320
	Borders	0.9456	0.7548	0.6043	0.9706	0.6699	0.9475
	Color (Grayscale)	0.9424	0.7216	0.5788	0.9784	0.6424	0.9432
	Color (Mask)	0.8502	0.6073	0.1136	0.9206	0.1914	0.9154
	Diameter	0.9553	0.7688	0.6513	0.9811	0.7052	0.9520

Table 3. Experimental results obtained by training (and testing) the models on the input configurations described in Sect. 4. Each of the *Experiment* correspond to a training performed on the corresponding debased ISIC19-20 dataset and tested on the debased private dataset. Threshold is set to 0.5.

Model	Experiment	AUC ROC	Precision	Recall (Sensitivity)	Specificity	F1-Score	Accuracy
EfficientNet-B3	Original	0.7983	0.5299	0.5038	0.9104	0.5165	0.8425
	Asymmetry	0.7693	0.5553	0.4025	0.9354	0.4667	0.8465
	Borders	0.7896	0.5261	0.4992	0.9099	0.5123	0.8413
	Color (Grayscale)	0.7673	0.4607	0.4540	0.8935	0.4573	0.8201
	Color (Mask)	0.7032	0.6017	0.0322	0.9957	0.0612	0.8349
	Diameter	0.8099	0.5597	0.5168	0.9185	0.5374	0.8515
ResNet-152	Original	0.7872	0.4774	0.5542	0.8772	0.5129	0.8229
	Asymmetry	0.7340	0.5279	0.3176	0.9416	0.3966	0.8351
	Borders	0.7559	0.4498	0.4921	0.8762	0.4700	0.8107
	Color (Grayscale)	0.6860	0.3565	0.4411	0.8389	0.3943	0.7719
	Color (Mask)	0.6881	0.5243	0.1187	0.8436	0.1936	0.8313
	Diameter	0.7660	0.4121	0.5424	0.8409	0.4684	0.7899

Networks are trained for 20 epochs and produce 9 class probabilities as output, among which only the melanoma class is considered.

Given the unavailability of ISIC test set labels, we expanded our evaluation metrics by partitioning the validation set of ISIC19-20 to create an "internal" test set: the resulting dataset counts 46 379 training images, 1 159 for validation, and 10 426 images for testing. The *private dataset* is employed for testing classification performance as well. These datasets, modified as outlined in Sect. 4



(a) Original (b) Borders (c) Color (d) Diameter

Fig. 5. Example of failure cases due to wrongly generated masks.

facilitated a broader analysis across five variant datasets, against which designated architectures are trained and tested. Table 2 and Table 3 report results obtained by training the model on the debased ISIC19-20 datasets and testing on the ISIC19-20 "internal" test set and on the private dataset, respectively.

6 Discussion

The efficacy of our ABCDE feature-concealment methods, despite occasional inaccuracies in segmentation mask generation (Fig. 5), underscores their ability to divert neural networks' focus from compromised features towards other ones, as demonstrated in most test scenarios (Fig. 3).

In particular, in Fig. 3c the neural network trained to classify images with hidden borders makes a prediction focusing on the section of the lesion with the most variance of color intensity. The same patch is equally important for the network fed with grayscale images (Fig. 3d), whereas Fig. 3a shows that mole borders are of great interest for the "standard" model.

On the other hand, CNN asked to make a prediction based solely on the segmentation mask strictly focuses on the sections of the foreground with higher concavity, which is roughly the only *valuable* piece of information about the lesion to be found in the extremely degraded input.

As suggested by the high classification performance, neural networks autonomously learn to extract features for melanoma detection. The accuracy obtained when single important features are missing is very close to the "reference" values, meaning that the other image features are enough to produce a satisfying classification prediction. Experimental results alone are clearly not enough to distinguish whether such features are biases or notoriously important elements for melanoma detection. For this reason, in our work, we also rely on the clinically validated Grad-CAM analysis (Fig. 3).

The discriminative power of CNNs is also confirmed on the private dataset. CNN performance tends to drop when the source domain (training data) and the target domain (test data) come from distinct origins, even for extremely simple tasks such as handwritten digit recognition, where classification accuracy

across separate datasets can be decreased by up to 40% [49]. A performance drop can be due to a large number of reasons (biases), such as different lighting settings, resolution, image quality, human-introduced artifacts, subject centering, and image acquisition devices [36,52].

Notably, this is also confirmed by our experiments, identifying that model generalization abilities are satisfactory (AUC performance is higher than those obtained by expert dermatologists [7]), but certainly require fine-tuning the models on the real case scenario they have to be employed, thus ensuring an adequate level of Precision and Recall.

About AUC. The Area Under the Receiver Operating Characteristic curve (AUC) is a well-known metric designed to evaluate the diagnostic capabilities of binary classifiers. It is the official metric of the ISIC2020 challenge, and it offers the advantage of not needing a fixed threshold, thus supplying one less parameter to "overfit" proposed algorithms on the official test set. However, real clinical applications require a threshold to be set and a class prediction to be given; evaluating experimental results uniquely using the AUC metric can be misleading. To put results into context, we further discuss the performance of the CNN trained to classify skin lesion binary masks (*i.e.,* debased dataset obtained removing colors). Focusing on the EfficientNet-B3 results in Table 2, the fifth line shows that the investigated network yields an AUC of 0.8017 when tested on the subset of public images (ISIC19-20) used as an "internal" test set. This is the area under the blue curve in Fig. 6. When following this strategy, we obtain a tool with a sensitivity of 0.0656, and a specificity of 0.9972, which means that the model can correctly recognize only 6.5% of the melanoma cases, but successfully identifies 99.7% of the not-melanoma cases. Clearly, this is not the positive result that an AUC of 0.8017 might suggest.

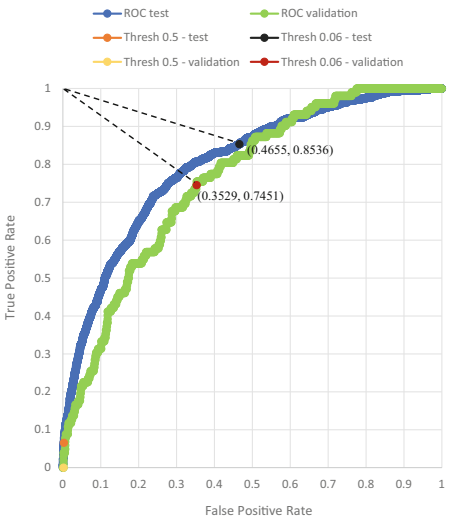


Fig. 6. ROC curves for the *Mask* experiment on the ISIC19-20 "internal" validation and test sets (Table 2 and Table 5 of the paper). The threshold value that minimizes the distance from the (0;1) are highlighted in both ROC curves. Moreover, the points corresponding to the 0.5 threshold are highlighted in the two curves.

As a matter of fact, the assumption that 0.5 is an appropriate threshold when dealing with neural networks is not correct [40], as shown in Fig. 6. Alternatively, the threshold can be set by studying the ROC curve obtained on the validation set (green curve in Fig. 6), and choosing the value in the graph closer to point

Table 4. Experimental results obtained by training (and testing) the models on the input configurations described in Sect. 4. Each of the *Experiment* correspond to a training performed on the corresponding debased ISIC19-20 dataset and tested on the debased ISIC19-20 "internal" test set using a specific threshold calculated as the value of the ROC curve which minimizes the distance from (0;1) on the validation set.

Model	Experiment	AUC ROC	Precision	Recall (Sensitivity)	Specificity	F1-Score	Accuracy
EfficientNet-B3	Original	0.9671	0.3163	0.9519	0.8020	0.4748	0.8152
	Asymmetry	0.9448	0.2527	0.9628	0.7260	0.4003	0.7468
	Borders	0.9605	0.2218	0.9858	0.6672	0.3621	0.6952
	Color (Grayscale)	0.9559	0.2590	0.9628	0.7349	0.4082	0.7549
	Color (Mask)	0.8017	0.1500	0.8536	0.5345	0.2551	0.5625
	Diameter	0.9724	0.2419	0.9803	0.7044	0.3881	0.7287

Table 5. Experimental results obtained by training (and testing) the models on the input configurations described in Sect. 4. Each of the *Experiment* correspond to a training performed on the corresponding debased ISIC19-20 dataset and tested on the private test set using a specific threshold calculated as the value of the ROC curve which minimizes the distance from (0;1) on the validation set.

Model	Experiment	AUC ROC	Precision	Recall (Sensitivity)	Specificity	F1-Score	Accuracy
EfficientNet-B3	Original	0.7983	0.2578	0.8570	0.5055	0.3963	0.5642
	Asymmetry	0.7693	0.2140	0.9017	0.3365	0.3460	0.4308
	Borders	0.7896	0.2012	0.9300	0.2600	0.3308	0.3718
	Color (Grayscale)	0.7673	0.2291	0.8844	0.4038	0.3639	0.4840
	Color (Mask)	0.7032	0.2421	0.7604	0.5229	0.3672	0.5626
	Diameter	0.8099	0.2150	0.9263	0.3221	0.3489	0.4230

$(0; 1)$, *i.e.,* the value that maximizes *True Positive Rate* while minimizing *False Positive Rate*. In this particular case, the desired rate is $\approx$0.06, and by employing this same threshold on the EfficientNet-B3 CNN outputs over the test set, we obtain a binary classifier with a sensitivity of 0.8536 and a specificity of 0.5345. Table 4, Table 5, and Table 6 present the results obtained by setting the prediction threshold following the described steps, always making use of the validation set. Regardless of how thresholds are set, it is clear that high AUC values do not always correspond to satisfying discriminative capabilities.

Table 6. Experimental results using foreground densities obtained from segmentation masks, bounding boxes, and bounding boxes that cover at least 70% of the image as melanoma probability on the ISIC19-20 "internal" test set and on the private dataset.

Dataset	Experiment	AUC ROC	Precision	Recall (Sensitivity)	Specificity	F1-Score	Acc.
ISIC19-20 "Internal" test set	Segm. Mask	0.7215	0.1483	0.7388	0.5917	0.2470	0.6046
	B. Box	0.7154	0.1483	0.7202	0.6019	0.2459	0.6123
	B. Box 70%	0.6220	0.1830	0.3989	0.8286	0.2509	0.7909
Private dataset	Segm. Mask	0.6980	0.2856	0.5898	0.7043	0.3848	0.6852
	B. Box	0.6919	0.2573	0.6589	0.6190	0.3701	0.6256
	B. Box 70%	0.6517	0.3328	0.4735	0.8098	0.3909	0.7536

Finally Identifying the "Bias" in Dermoscopic Datasets. Contrasting with Bissoto et al.'s findings [4] where CNNs performed well despite significant lesion occlusion, our analysis suggests that lesion size can be inferred from the foreground-background ratio and significantly influences predictions. While Bissoto et al. observed high AUCs (0.712) with major lesion coverage ($\geq$ 70%), our evaluations posit that networks might rely on lesion dimensions rather than intricate pixel patterns unrelated to the mole. This is substantiated by our *Segmentation Mask* and *Bounding Box* experiments (Table 6), where AUCs correlate strongly with lesion area metrics, even without deep learning models. This experiment has been pushed further by making predictions based only on lesion bounding box (and not segmentation mask) dimensions and, finally, by setting the foreground-background ratios as $\geq$ 70%. Results obtained are reported in the aforementioned table with the name of *Bounding Box* and *Bounding Box 70%*. Finally, histograms in Fig. 4 show that the probability of a lesion being malignant grows with its size within a dermoscopic image. Intuitively, a mole that exceeds the borders or gets very close to them is not necessarily larger than others, but it is more likely to be malignant. This characteristic might be more related to the complexity of including a whole malignant lesion when acquiring dermoscopic images [13,43], than to the diameter itself. Nevertheless, this feature is strongly related to the nature of dermoscopic images, and experiments provided in [4] are insufficient to prove the presence of biases in the ISIC dataset.

7 Conclusion

In this work, we explored the correlation between automatic skin lesion classification and the ABCDE rule. This was done by gradually removing important visual information from CNN inputs and analyzing performance changes. Experimental results show that neural networks autonomously learn to extract features that are notoriously important for melanoma detection, but also prove that their performance is still satisfying when some of these features are removed. Our experiments provide *no proof* that this is related to dataset biases: instead, the

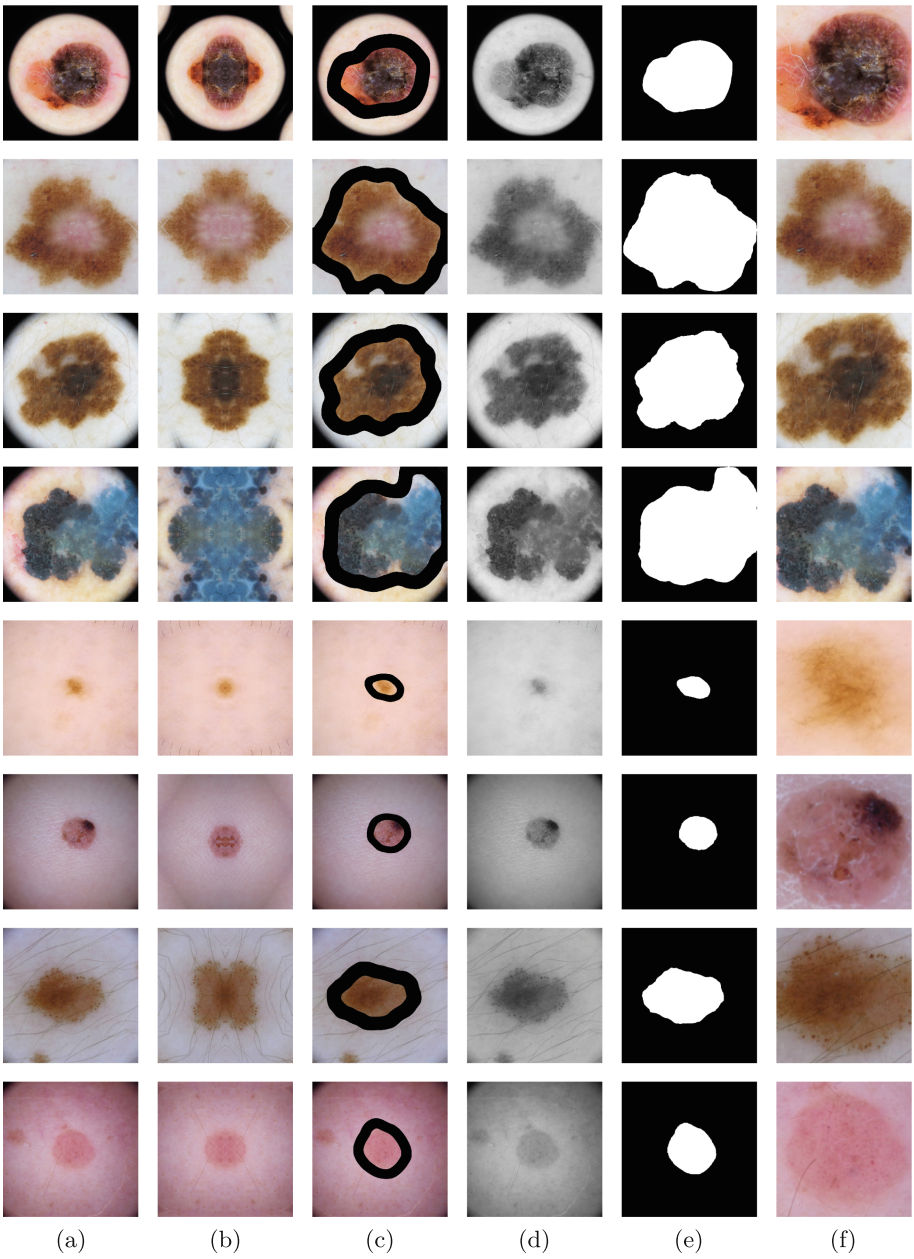


Fig. 7. Skin lesion image samples obtained from ISIC dataset after debasing different ABCDE properties. Columns from left to right: (a) Original, (b) Asymmetry, (c) Borders, (d) Color - Grayscale, (e) Color - Mask, (f) Diameter. The first half rows depict melanomas, while the others are generated from benign lesions.

remaining information can be enough to achieve satisfying or even good classification accuracy. As pointed out by different authors [44,47], the interpretation of GradCAM's saliency maps may be subjective to reader biases and cannot be used to draw general conclusions about network behavior. However, combined with the quantitative evaluation discussed and showcased in this paper, they contribute to our final conclusion.

In particular, the proposed paper experimentally proved that the foreground-background ratio is strongly related to the malignancy probability of a skin lesion. The reasoning behind this might be related to the well-known *diameter* characteristic from the ABCDE rule, but also to the fact that capturing the entire malignant mole in a dermoscopic image is usually not trivial given its dimensions, the non-clearly defined borders, and the irregular shapes that characterize cancerous skin lesions [8,11,13,43]. Nevertheless, foreground-background ratio is a valuable dermoscopic property. We cannot conclude that "there are no biases in the ISIC dataset", but we can certainly state that literature claims of strong biases affecting the ISIC dataset are supported by an inconsistent experimental analysis.

Finally, testing model performance on a totally distinct private dataset, with no possible intersection with samples employed during the training phase, demonstrated that, despite intra-datasets biases (if any), state-of-the-art algorithms preserve satisfactory performance: still higher than those obtained by expert dermatologists [7], but with lower Precision and Recall.

Acknowledgements. This work was supported by the University of Modena and Reggio Emilia and Fondazione di Modena, through the FAR 2023 and FARD-2023 funds (Fondo di Ateneo per la Ricerca).

References

1. Abayomi-Alli, O.O., Damasevicius, R., Misra, S., Maskeliunas, R., Abayomi-Alli, A.: Malignant skin melanoma detection using image augmentation by oversamplingin nonlinear lower-dimensional embedding manifold. Turkish J. Electr. Eng. Comput. Sci. **29**(8) (2021)
2. Allegretti, S., Bolelli, F., Pollastri, F., Longhitano, S., Pellacani, G., Grana, C.: Supporting skin lesion diagnosis with content-based image retrieval. In: 2020 25th International Conference on Pattern Recognition (ICPR). IEEE (2021)
3. Barata, C., Celebi, M.E., Marques, J.S.: Explainable skin lesion diagnosis using taxonomies. Pattern Recogn. **110** (2020)
4. Bissoto, A., Fornaciali, M., Valle, E., Avila, S.: (De)constructing bias on skin lesion datasets. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (2019)
5. Bolelli, F., Baraldi, L., Grana, C.: A hierarchical quasi-recurrent approach to video captioning. In: 2018 IEEE International Conference on Image Processing, Applications and Systems (IPAS) (2018)
6. Bray, F., Ferlay, J., Soerjomataram, I., Siegel, R.L., Torre, L.A., Jemal, A.: Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA: Cancer J. Clin. **68**(6) (2018)

7. Brinker, T.J., et al.: Comparing artificial intelligence algorithms to 157 German dermatologists: the melanoma classification benchmark. Eur. J. Cancer **111** (2019)
8. Brú, A., Albertos, S., Subiza, J.L., García-Asenjo, J.L., Brú, I.: The universal dynamics of tumor growth. Biophys. J. **85**(5) (2003)
9. Canalini, L., Pollastri, F., Bolelli, F., Cancilla, M., Allegretti, S., Grana, C.: Skin lesion segmentation ensemble with diverse training strategies. In: Vento, M., Percannella, G. (eds.) CAIP 2019. LNCS, vol. 11678, pp. 89–101. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-29888-3_8
10. Cancilla, M., et al.: The DeepHealth toolkit: a unified framework to boost biomedical applications. In: 2020 25th International Conference on Pattern Recognition (ICPR). IEEE (2021)
11. Capdehourat, G., Corez, A., Bazzano, A., Alonso, R., Musé, P.: Toward a combined tool to assist dermatologists in melanoma detection from dermoscopic images of pigmented skin lesions. Pattern Recogn. Lett. **32**(16) (2011)
12. Chen, L.C., Zhu, Y., Papandreou, G., Schroff, F., Adam, H.: Encoder-decoder with atrous separable convolution for semantic image segmentation. In: Proceedings of the European Conference on Computer Vision (ECCV) (2018)
13. Claridge, E., Hall, P., Keefe, M., Allen, J.: Shape analysis for classification of malignant melanoma. J. Biomed. Eng. **14**(3) (1992)
14. 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)
15. Di Biasi, L., De Marco, F., Auriemma Citarella, A., Castrillón-Santana, M., Barra, P., Tortora, G.: Refactoring and performance analysis of the main CNN architectures: using false negative rate minimization to solve the clinical images melanoma detection problem. BMC Bioinform. **24** (2023)
16. Esteva, A., et al.: Dermatologist-level classification of skin cancer with deep neural networks. Nature **542**(7639) (2017)
17. Fujisawa, Y., et al.: Deep-learning-based, computer-aided classifier developed with a small dataset of clinical images surpasses board-certified dermatologists in skin tumour diagnosis. Br. J. Dermatol. **180**(2) (2019)
18. Gaudy-Marqueste, C., et al.: Ugly duckling sign as a major factor of efficiency in melanoma detection. JAMA Dermatol. **153**(4) (2017)
19. Geirhos, R., et al.: Shortcut learning in deep neural networks. Nat. Mach. Intell. **2**(11) (2020)
20. Geirhos, R., Rubisch, P., Michaelis, C., Bethge, M., Wichmann, F.A., Brendel, W.: ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness. In: International Conference on Learning Representations (2019)
21. Ha, Q., Liu, B., Liu, F.: Identifying melanoma images using efficientnet ensemble: winning solution to the SIIM-ISIC melanoma classification challenge. arXiv preprint arXiv:2010.05351 (2020)
22. Hassija, V., et al.: Interpreting black-box models: a review on explainable artificial intelligence. Cogn. Comput. **16**(1), 45–74 (2024)
23. 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) (2016)
24. Hernández-Pérez, C., et al.: BCN20000: dermoscopic lesions in the wild. Sci. Data (2024)

25. Hinton, G., et al.: Deep neural networks for acoustic modeling in speech recognition: the shared views of four research groups. IEEE Signal Process. Mag. **29**(6) (2012)
26. Kadry, S., Taniar, D., Damaševičius, R., Rajinikanth, V., Lawal, I.A.: Extraction of abnormal skin lesion from dermoscopy image using VGG-SegNet. In: 2021 Seventh International Conference on Bio Signals, Images, and Instrumentation (ICBSII) (2021)
27. Kingma, D., Ba, J.: Adam: a method for stochastic optimization. In: International Conference on Learning Representations (ICLR), San Diego, CA, USA (2015)
28. Lattoofi, N.F., et al.: Melanoma skin cancer detection based on ABCD rule. In: 2019 First International Conference of Computer and Applied Sciences (CAS). IEEE (2019)
29. Lynn, N.C., Kyu, Z.M.: Segmentation and classification of skin cancer melanoma from skin lesion images. In: 2017 18th International Conference on Parallel and Distributed Computing, Applications and Technologies (PDCAT). IEEE (2017)
30. Maguolo, G., Nanni, L.: A critic evaluation of methods for COVID-19 automatic detection from X-ray images. arXiv preprint arXiv:2004.12823 (2020)
31. Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., Galstyan, A.: A survey on bias and fairness in machine learning. ACM Comput. Surv. (CSUR) **54**(6) (2021)
32. Mondal, B., Das, N., Santosh, K., Nasipuri, M.: Improved skin disease classification using generative adversarial network. In: 2020 IEEE 33rd International Symposium on Computer-Based Medical Systems (CBMS). IEEE (2020)
33. Nawaz, M., et al.: Melanoma segmentation: a framework of improved DenseNet77 and UNET convolutional neural network. Int. J. Imaging Syst. Technol. (2022)
34. Olteanu, A., Castillo, C., Diaz, F., Kıcıman, E.: Social data: biases, methodological pitfalls, and ethical boundaries. Front. Big Data **2** (2019)
35. Pellacani, G., Grana, C., Seidenari, S.: Algorithmic reproduction of asymmetry and border cut-off parameters according to the ABCD rule for dermoscopy. J. Eur. Acad. Dermatol. Venereol. **20**(10) (2006)
36. Perone, C.S., Ballester, P., Barros, R.C., Cohen-Adad, J.: Unsupervised domain adaptation for medical imaging segmentation with self-ensembling. NeuroImage **194** (2019)
37. Pollastri, F., et al.: Confidence calibration for deep renal biopsy immunofluorescence image classification. In: 2020 25th International Conference on Pattern Recognition (ICPR). IEEE (2021)
38. Pollastri, F., et al.: A deep analysis on high resolution dermoscopic image classification. IET Comput. Vis. (2021)
39. Rigel, D.S., Russak, J., Friedman, R.: The evolution of melanoma diagnosis: 25 years beyond the ABCDs. CA: Cancer J. Clin. **60**(5) (2010)
40. Roberts, M., et al.: Common pitfalls and recommendations for using machine learning to detect and prognosticate for COVID-19 using chest radiographs and CT scans. Nat. Mach. Intell. **3**(3) (2021)
41. Rotemberg, V., et al.: A patient-centric dataset of images and metadata for identifying melanomas using clinical context. Sci. Data **8**(1) (2021)
42. 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: 2017 IEEE International Conference on Computer Vision (ICCV) (2017)
43. Senan, E.M., Jadhav, M.E.: Analysis of dermoscopy images by using ABCD rule for early detection of skin cancer. Glob. Transit. Proc. **2**(1) (2021)

44. Srinivas, S., Fleuret, F.: Rethinking the role of gradient-based attribution methods for model interpretability. In: International Conference on Learning Representations (2021)
45. Suresh, H., Guttag, J.: A Framework for understanding sources of harm throughout the machine learning life cycle. In: Equity and Access in Algorithms, Mechanisms, and Optimization. Association for Computing Machinery (2021)
46. Tan, M., Le, Q.: EfficientNet: rethinking model scaling for convolutional neural networks. In: International Conference on Machine Learning (2019)
47. Viviano, J.D., Simpson, B., Dutil, F., Bengio, Y., Cohen, J.P.: Saliency is a possible red herring when diagnosing poor generalization. In: International Conference on Learning Representations (2021)
48. Wang, H., et al.: Score-CAM: score-weighted visual explanations for convolutional neural networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (2020)
49. Wang, M., Deng, W.: Deep visual domain adaptation: a survey. Neurocomputing **312** (2018)
50. Weber, J.: True duplicates in ISIC 2020 dataset (2020). https://www.kaggle.com/c/siim-isic-melanoma-classification/discussion/161943
51. Wu, Y., Chen, B., Zeng, A., Pan, D., Wang, R., Zhao, S.: Skin cancer classification with deep learning: a systematic review. Front. Oncol. (2022)
52. You, K., Long, M., Cao, Z., Wang, J., Jordan, M.I.: Universal domain adaptation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2019)
53. Zhou, B., Bau, D., Oliva, A., Torralba, A.: Interpreting deep visual representations via network dissection. IEEE Trans. Pattern Anal. Mach. Intell. **41**(9) (2019)

Breast Cancer Segmentation Using UNet and Global Convolutional Networks

Anand Thyagachandran(✉) and Yeruru Asrar Ahmed

Department of Computer Science and Engineering, Indian Institute of Technology Madras, Chennai, India
{tanand,asrar}@cse.iitm.ac.in

Abstract. Breast ultrasound (BUS) imaging techniques have become efficient tools for cancer diagnosis. Convolutional neural network (CNN) based encoder-decoder architectures have been widely used for the automated segmentation of tumours in BUS images, assisting in breast cancer diagnoses. However, these models have limitations in capturing long-range dependencies. To overcome this limitation, various deep learning techniques, such as atrous convolution, attention mechanisms, and transformer encoder-based models, have been introduced to capture long-range dependencies in feature maps, improving segmentation accuracy by considering larger receptive fields and global context. As modelling techniques evolve, there is a shift towards more complex and intricate designs. This study proposes a simple yet effective model that combines UNet and Global Convolutional Network (GCN) architectures for breast lesion segmentation. By leveraging the GCN block, our model captures broader receptive fields with a simpler design strategy. We have demonstrated the efficacy of our approach through various experiments, including kernel size analysis, model component evaluation, and data preprocessing assessment. The proposed model has been evaluated using four-fold cross-validation with BUSI and Dataset-B datasets. Additionally, models trained on both datasets have been validated with a blind test dataset, where our model demonstrates better performance compared to state-of-the-art methods, achieving a 4.9% and 6.7% improvement in Intersection over Union (IoU) score, respectively. The robustness analysis and external validation experiments underscore the superior generalization performance of our model in breast lesion segmentation tasks.

Keywords: Breast cancer · segmentation · ultrasound imaging

1 Introduction

Breast cancer is a significant global health challenge, especially among women, and it has high mortality rates [1]. Early detection of symptoms is crucial for effective treatment. Breast ultrasound (BUS) imaging techniques have emerged as efficient tools for cancer diagnosis. They are cost-effective, non-invasive, provide real-time results, and do not involve ionizing radiation [2]. Breast cancer is

A. Antonacopoulos et al. (Eds.): ICPR 2024, LNCS 15313, pp. 112–126, 2025.
https://doi.org/10.1007/978-3-031-78201-5_8

classified into benign and malignant breast lesions, where benign breast lesions pose no threat to health, while malignant breast lesions are cancerous growths that can spread throughout the body. The diagnosis of breast cancer involves detection, segmentation, and classification stages. This study mainly focuses on segmenting the lesion regions from BUS images to aid in the diagnosis of breast cancer.

The medical field has seen significant advancements in the automation of medical image segmentation, providing valuable assistance to clinicians in quantitative pathological assessment and diagnosis. Segmentation methods can broadly divided into semi-automated and fully automated procedures based on the manual intervention to fine-tune the breast lesion regions in BUS images [3,4]. A comprehensive review of earlier methods is available in Xian *et al.* [5]. Fully automatic methods, exemplified by recent works such as Chen *et al.* [6] and Yan *et al.* [7], eliminate the need for user intervention. These methods primarily leverage convolutional operations at each layer to extract local image features from neighbouring pixels, enabling these models to predict the semantics of objects in medical images. Among them, fully convolutional encoder-decoder-based models like UNet [8] are extensively utilized. The UNet architecture comprises encoder and decoder layers to extract features for predicting maps at image resolution. Additionally, it uses skip connections between the encoder and decoder layers to preserve the spatial structure, enabling precise object localization. The effectiveness of the UNet architecture is evident in its state-of-the-art performances, particularly with small medical image datasets, owing to its compact parameterization and encoder-decoder design. Variants of the UNet model are widely adopted in biomedical image segmentation tasks [9–13].

Almajalid *et al.* introduced UNet for breast lesion segmentation from BUS images [14]. Later, various UNet variants were proposed to refine segmentation accuracy. These variants can be broadly categorized into four classes: multiscale UNet [6,15–17], attention-based UNet [7,18,19], deep supervised UNet [20–22], and multi-module hybrid UNet [23,24]. Multiscale UNet models utilize diverse convolutional kernel sizes to capture context information across different receptive fields. Attention-based UNet models aim to capture global context information for improved segmentation, including hybrid dilated convolution-based attention UNet [7] and channel attention module [19]. Deep supervised UNet ensures each stage contributes to the loss function, enforcing feature learning proximity to the ground truth at every stage of the model. Multi-module hybrid UNet architectures integrate disparate, independent modules-such as Tversky loss functions [20], residual inception depth-wise separable convolutions, and hybrid pooling strategies (combining max pooling and spectral pooling)-alongside cross-spatial attention filters [23] to further refine segmentation predictions.

The convolution based encoder-decoder models encounter challenges associated with capturing long-range dependencies between pixels in the feature maps. The convolutional operation often leads to inductive biases [25], limiting architectures' ability to model long-range feature dependencies effectively. Two

strategies are commonly employed to address the limitations, such as enlarging the receptive field [26–31] and incorporating attention mechanisms [32–34]. Atrous convolution operations [25] are utilized to insert holes into convolution kernels, preserving resolution and enlarging the receptive field. However, relying solely on atrous convolution operations may not fully address challenges posed by surrounding tissues and indistinct boundaries [15]. Attention mechanisms have also been integrated to exploit long-range dependencies in CNNs [11,19,20,34]. These mechanisms enhance models' capacity to capture intricate details and disregard irrelevant features by dynamically focusing on relevant regions within the input image [11]. As modelling choices evolve, there is a shift towards more complex and intricate designs. Several studies have utilized a combination of atrous convolution and attention mechanisms [6,7,24] to enhance tumour segmentation in Breast Ultrasound (BUS) images, categorizing them as hybrid models.

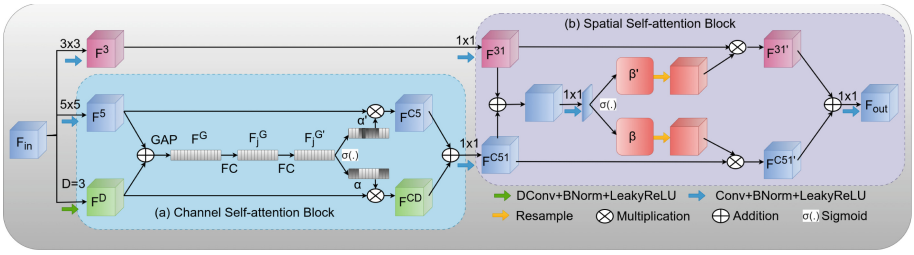


Fig. 1. The HAAM block architecture

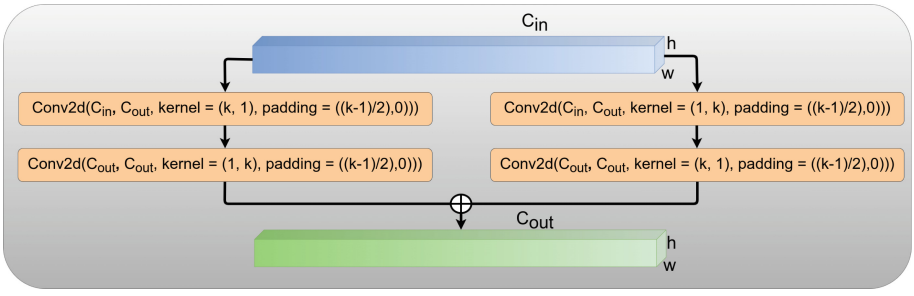


Fig. 2. The GCN block architecture

A recent development in hybrid models for breast cancer segmentation is the Adaptive Attention UNet (AAUNet) [6]. In this model, Chen *et al.* replaced the conventional convolution layers in the UNet encoder and decoder blocks with a Hybrid Adaptive Attention Module (HAAM) [6]. The intricate design of

the AAUNet model has demonstrated superior performance compared to state-of-the-art semantic segmentation models with BUS images. The HAAM block, a key component of AAUNet, effectively captures a larger receptive field by employing multiple convolutions with varying receptive fields. The channel and spatial attention modules use these output features to enhance the segmentation accuracy. The HAAM block diagram is shown in Fig. 1. In contrast, our proposed approach presents a simpler architecture, leveraging the UNet framework augmented with a Global Convolutional Network (GCN) [35]. The GCN block primarily employs separable large filters to capture extensive receptive fields while minimizing the number of parameters compared to standard convolutions. The GCN has a symmetric structure to capture broader and better receptive fields and is shown in Fig. 2. By integrating large receptive field information, the GCN blocks allow superior segmentation predictions [35]. The combination of UNet, Global Convolution Network (GCN), and Boundary Refinement (BR) Module has been previously proposed for segmenting tongue medical images, as demonstrated by [36]. GCN and BR blocks are employed to reduce the gap between localization and classification. Moreover, the GCN block is incorporated randomly in the UNet model. In contrast, the purpose of the GCN block in our model is to increase the receptive field.

The remainder of this paper is organised as follows: Sect. 2 discusses the proposed segmentation model for BUS images. Section 3 outlines the datasets and evaluation metrics utilized in the study. Experimental results and inferences are presented in Sect. 4. Finally, Sect. 5 outlines the future scope and conclusions drawn from the work.

2 Methodology

Our proposed model presents a UNet-based architecture including a Global Convolutional Network (GCN) block. Utilizing the UNet as its foundation, the proposed architecture comprises five encoder layers, five decoder layers, and skip connections between encoder and decoder for preserving spatial structure. GCN block is used to expand its receptive field, empowering it to extract contextual information effectively. The GCN block integrates into the skip connection structure through empirical analysis, enhancing its capacity to capture spatial relationships and semantic context in medical imaging data. This simple approach utilises the strengths of GCN to improve segmentation performance, particularly in tasks requiring capturing spatial and semantic relations. The proposed model architecture is shown in Fig. 3.

2.1 UNet

The UNet, introduced by Ronneberger *et al.* in 2015 [8], is an encoder-decoder architecture based on fully convolutional neural networks. In the UNet, the encoder layer initially captures high-frequency features and gradually refines them for semantic extraction across subsequent encoder layers. Multiple encoder

layers with max-pooling downsample the image to low-resolution feature maps, which are then passed to a bottleneck layer. These features are upsampled using decoder layers during the decoding process, and features from corresponding skip connections are incorporated. Skip connections help preserve the spatial structure, while the upsampled features from decoder layers capture more semantic information, facilitating precise identification of regions of interest within medical images. Each encoder and decoder layer consists of two convolution layers (3×3 kernel), followed by Instance Normalization [37] and LeakyReLU activation functions.

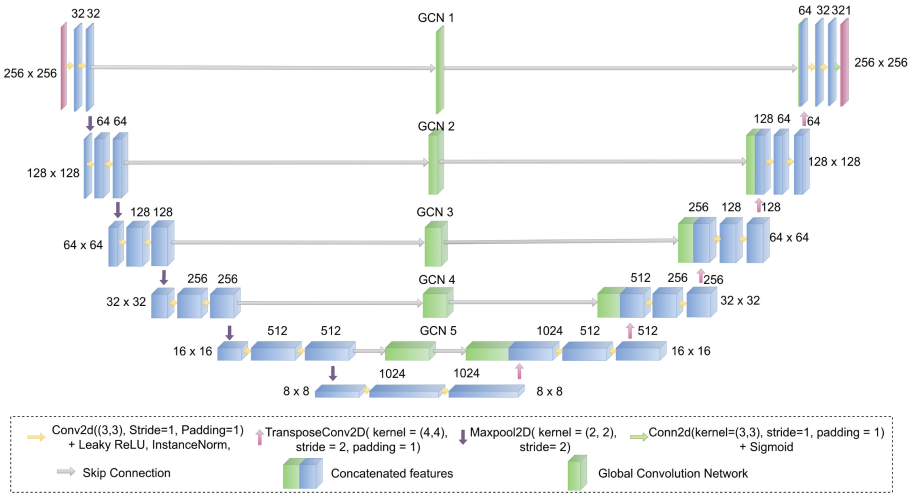


Fig. 3. The proposed UNet-GCN architecture

2.2 Global Convolutional Network (GCN)

Medical image segmentation models typically use 3×3 kernels in convolutional layers [8,13,14], primarily capturing local information and limiting larger receptive fields in initial layers. While atrous convolution addresses this limitation by employing dilated convolutions [6], it often provides only large, sparse receptive fields [12]. A simpler approach would be using larger kernels in convolution, which increases the receptive field and aids in handling significant variations in lesion transformations in BUS images [15]. However, using large kernels increases exponentially the number of parameters and GPU memory usage in each convolution layer. GCN block [35] captures a larger receptive field with linear growth in parameters and can easily be incorporated into existing architectures.

GCN approximates $k\times k$ convolutions using four low-rank convolutions in two parallel branches. Each branch consists of two low-rank convolution kernel sizes of $k \times 1$ followed by $1 \times k$ and vice versa. The dual branch gives equal precedence

to the horizontal and vertical kernels to capture the information. Traditional $k \times k$ convolutions require k^2 parameters, whereas parameters increase linearly in GCN and need only $4k$ parameters. GCN replicates dense connections to the input feature map within the receptive field of $k \times k$, which helps handle large variations of transformations. GCN approach helps to increase the receptive field in the early stages while reducing overall parameter growth in the model. Self-attention cannot be used in early layers to capture larger contexts due to the exponential increase in GPU memory, a challenge mitigated by GCN blocks. The optimal position of GCN within the backbone network and the kernel sizes of the GCN block are determined empirically, as detailed in Sect. 4.1, and Sect. 4.2. The block diagram of GCN is shown in Fig. 2.

2.3 Preprocessing BUS Images

Ultrasound images often suffer from low signal-to-noise ratio (SNR) and various artefacts like speckle noise, reverberations, and acoustic shadowing, which degrade image quality [38]. Image preprocessing techniques are commonly employed to enhance BUS images and their quality [38]. Image preprocessing methods like contrast enhancement, brightness adjustment, Gaussian blurring and histogram equalization are employed in our work. Gaussian blurring removes the high-frequency noise and preserves the structure and edges in the image. Histogram equalization [39] redistributes the intensity values across the histogram to enhance the quality of the image. This process effectively stretches the intensity levels, making the image appear more visually appealing with improved contrast and detail. These preprocessing methods are integrated into the data augmentation, as detailed in Table 1. Specifically, we employ six image transformations, with three randomly selected transformations applied to each image in the batch during training. Such augmentation strategies have been demonstrated to enhance model performance in semantic segmentation tasks significantly [40].

Table 1. Different image preprocessing methods employed in the Data augmentation process.

Image Preprocessing	Description
Identity	Returns the original image.
Gaussian blur	Blur the image with a Gaussian kernel.
Equalize	Histogram Equalization
Contrast	Adjusts the contrast of the image by [0.05, 0.95].
Brightness	Adjusts the brightness of the image by [0.05, 0.95]
Random Flip	Horizontally flips the image with a probability of 0.5.

2.4 Model Settings

All the images are reshaped into 256 × 256 pixels before being input into the proposed model. During model training, a stratified batch (equal number of images from the breast lesion class and normal class) is used to avoid bias in the class imbalance dataset [41]. Binary cross entropy [42] is employed as the loss function.

$$\mathcal{L}_{BCE} = -\sum_{(i,j)} GT(i,j) * \log(PD(i,j)) + (1 - GT(i,j)) * (1 - \log(PD(i,j)))$$

where $GT(i,j) \in [0,1]$ denotes the ground-truth mask (i,j), $PD(i,j) \in [0,1]$ represents the predict masks. Adam optimizer with a learning rate of 0.0001 is used for model optimization. The model is trained for 100 epochs with a batch size of 16. The parameters of the model are optimized on a validation set. Balanced data from both normal and lesion classes are used for training to maintain fairness. The code is written in PyTorch [43], and all experiments are conducted using two GeForce GTX 1080 Titans with an overall 24 GB GPU memory.

3 Datasets and Evaluation Metrics

3.1 BUS Dataset

The BUSI dataset, collected at Baheya Hospital for Early Detection & Treatment of Women's Cancer, Cairo, Egypt, in 2018 [44], consists of 780 BUS images obtained from 600 patients aged 25 to 75 years. The dataset encompasses three distinct classes of BUS images: benign (487 images), malignant (210 images), and normal (133 images). Imaging data was captured using the LOGIQ E9 ultrasound and LOGIQ E9 Agile ultrasound systems. Following the acquisition, skilled radiologists preprocessed the images to delineate lesion regions and eliminate extraneous areas. Subsequently, the images were converted to PNG format for standardized analysis.

Dataset-B [45] consists of 163 images, including 110 benign images and 53 malignant images. This dataset was captured using the Siemens ACUSON Sequoia C512 system at the UDIAT Diagnostic Centre of the Parc Taulí Corporation, Sabadell, Spain. Additionally, the STU dataset [24] contains 42 BUS images and corresponding masks. These images were acquired using the GE Voluson E10 Ultrasound Diagnostic System at Shantou First Affiliated Hospital, Guangdong Province, China. While all images in the STU dataset depict lesions, they are not explicitly classified as benign or malignant. The STU dataset is an external validation (test) dataset for evaluating model performance.

3.2 Evaluation Metrics

Image segmentation evaluation metrics are helpful in assessing the effectiveness of segmentation models. Five widely recognized metrics are used in our

work: Intersection over Union (IoU), Dice similarity coefficient (Dice), Precision (Prec.), Sensitivity (Sen.), and Specificity (Spec.). IoU is also known as the Jaccard index, which estimates the ratio of the intersection area between the prediction and ground truth mask. The dice score is also referred to as the F1 score, which estimates the ratio of twice the overlap between the prediction with ground truth mask to the sum of their areas. IoU and DSC evaluate the spatial correspondence between the predicted and ground truth masks, with higher values indicating superior segmentation accuracy. Precision estimates the proportion of correctly classified lesion pixels to the total number of lesion pixels predicted in the prediction mask, while Sensitivity measures the proportion of correctly classified lesion pixels in the prediction mask to the ground truth mask. Moreover, Specificity assesses the proportion of correctly classified background pixels in the prediction mask to the ground truth mask.

4 Experimental Settings and Results

This section presents a series of experiments to evaluate the performance of the proposed and baseline models in breast cancer segmentation using BUS images. An ablation study is conducted to understand the importance of GCN components within the model architecture. We also investigate kernel size's impact on breast lesion segmentation in the GCN block. Another ablation study evaluates the effect of data preprocessing techniques on BUS medical image segmentation. We then present and discuss the segmentation results obtained with state-of-the-art models using the BUSI and Dataset-B. All experiments are conducted using four-fold cross-validation on the sorted dataset and employ internal shuffling for uniformity. Finally, we assess our proposed and baseline models' generalizability using the unseen (Test) STU dataset. The STU dataset consists of two classes, tumour and normal, and the trained models predict whether each pixel in the BUS image is normal or a tumour.

4.1 GCN Position

The GCN is an independent block used to capture larger receptive fields and can be easily integrated into the UNet architecture. We explore three variants: Model A, where the GCN block is within the skip connection; Model B, where it's placed between each encoder and decoder block; and Model C, where it replaces each convolution in both encoder and decoder blocks (except for the upsampling convolution). Table 2 shows the performance of these variants, with skip connections proving to be the optimal choice in terms of performance and is employed for further analysis.

Table 2. Segmentation results for GCN at different positions in the proposed network with BUSI dataset. Models A, B, and C are defined in Sect. 4.1.

Models	BUSI				
	IoU	Dice	Sensitivity	Precision	Specificity
Model A	61.05 ± 1.31	75.69 ± 1.00	72.13 ± 1.35	78.29 ± 2.98	98.28 ± 0.31
Model B	60.28 ± 1.55	75.07 ± 1.21	72.12 ± 1.39	78.41 ± 3.39	98.26 ± 0.40
Model C	59.96 ± 1.07	74.79 ± 0.83	70.98 ± 2.97	79.06 ± 3.47	97.36 ± 0.44

4.2 Kernel Size

Using larger kernels enables the model to have larger receptive fields, enhancing its ability to predict lesions effectively. In an ablation study, we tested three different kernel sizes (3, 5, and 7) for the k parameter in the GCN, ensuring uniformity across all GCN kernels. We opt for a maximum kernel size of 7, restricted by the smallest feature size of 8×8 within the proposed network. It is observed in Table 3 that the larger kernel is progressively improving the model's segmentation prediction.

Table 3. Segmentation results for different kernels used in the GCN block with BUSI dataset.

Kernel Size (k)	BUSI				
	IoU	Dice	Sensitivity	Precision	Specificity
7	61.05 ± 1.31	75.69 ± 1.00	72.13 ± 1.35	78.29 ± 2.98	98.28 ± 0.31
5	60.51 ± 0.81	75.26 ± 0.66	72.04 ± 3.09	79.08 ± 3.65	98.29 ± 0.49
3	59.44 ± 1.10	74.72 ± 0.84	72.37 ± 2.64	76.69 ± 1.53	98.06 ± 0.23

4.3 Data Augmentation

BUS images are characterised by noise and low quality, often exhibiting low contrast. We apply domain knowledge-based data augmentation methods to address these issues and enhance image perception to improve contrast and reduce noise. Rather than adding domain-based augmentation directly, random augmentation settings are used for a superior augmentation approach [40]. To verify the claim that such a data augmentation method improves the perception quality and aids in the prediction of maps by the network, we perform an ablation study involving data augmentation techniques. We train our model with and without augmentation approaches and report its results in Table 4. The model trained with data augmentation achieves superior Dice and IoU scores compared to methods that do not utilise augmentation.

Table 4. Segmentation results of the proposed model with and without data augmentation using BUSI dataset.

Data Augmentation	BUSI				
	IoU	Dice	Sensitivity	Precision	Specificity
✓	61.05 ± 1.31	75.69 ± 1.00	72.13 ± 1.35	78.29 ± 2.98	98.28 ± 0.31
✗	59.31 ± 0.59	74.28 ± 0.50	70.59 ± 1.29	78.53 ± 2.24	98.29 ± 0.22

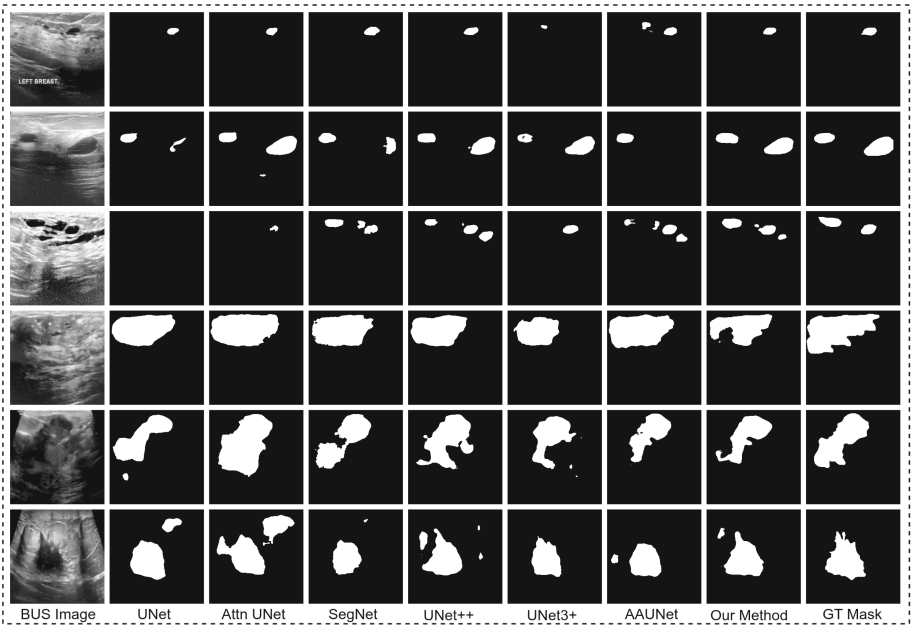


Fig. 4. The segmentation results of different methods on breast ultrasound images. The first column represents the input image. The remaining columns represent the corresponding mask predicted by the models.

Table 5. The cross-fold validation segmentation results for the baseline and proposed model. Sen, Prec, and Spec represent sensitivity, precision and specificity, respectively. NoP represents the number of parameters of the model, and the values are expressed in millions. FLOPS represents the Floating-point operations per second, and values are denoted in gigabits per second (Gbps).

Models	BUSI					Dataset-B					NoP	FLOPS
	IoU	Dice	Sen.	Prec.	Spec.	IoU	Dice	Sen.	Prec.	Spec.		
UNet	53.82 ± 2.59	69.75 ± 2.10	65.96 ± 4.67	74.07 ± 4.49	97.98 ± 0.57	60.83 ± 3.29	75.62 ± 2.62	68.32 ± 3.35	84.70 ± 5.85	99.37 ± 0.35	39	27.78
Attention UNet	57.08 ± 1.12	72.57 ± 0.90	70.89 ± 3.15	74.65 ± 3.99	97.87 ± 0.57	69.98 ± 2.68	82.29 ± 1.82	78.48 ± 5.03	86.96 ± 5.75	99.41 ± 0.28	34	66.69
UNet ++	57.14 ± 0.88	72.60 ± 2.21	69.15 ± 2.82	76.53 ± 3.59	98.16 ± 0.44	68.14 ± 2.17	80.99 ± 1.61	80.35 ± 5.75	82.25 ± 5.86	99.14 ± 0.37	47	199.85
UNet 3+	56.93 ± 0.95	72.43 ± 0.79	68.45 ± 2.09	77.27 ± 2.15	98.20 ± 0.29	69.86 ± 2.29	82.23 ± 1.53	78.02 ± 4.03	87.06 ± 4.01	99.39 ± 0.17	26	198.03
SegNet	57.55 ± 1.44	72.93 ± 1.18	68.04 ± 2.85	78.73 ± 2.47	98.34 ± 0.30	68.38 ± 2.54	81.20 ± 1.88	77.75 ± 2.97	85.23 ± 3.72	99.32 ± 0.20	29	40.82
AAUNet	59.90 ± 2.24	74.72 ± 1.76	69.62 ± 2.99	80.68 ± 5.59	98.52 ± 0.62	70.02 ± 2.83	82.34 ± 1.93	78.42 ± 3.76	86.70 ± 4.15	99.40 ± 0.26	43	85.33
Proposed Method	61.05 ± 1.31	75.69 ± 1.00	72.13 ± 1.35	78.29 ± 2.98	98.28 ± 0.31	72.11 ± 1.92	83.77 ± 1.29	83.96 ± 3.46	83.72 ± 2.82	99.02 ± 0.18	80	37.08

4.4 Comparison with State-of-the-Art Models

We have compared approaches like AAUNet, designed explicitly for breast lesion segmentation, with our proposed model performance. The other state-of-the-art medical segmentation methods include UNet [8], SegNet [46], Attention UNet [11], UNet++ [13], and UNet3+ [9] are also assessed. We have used the officially available repositories of these models to reproduce the results on the BUSI and other datasets. All models are performed four-fold cross-validation with data augmentation, and the results are shown in Table 5. Models were trained separately with four-fold cross-validation with datasets BUSI and Dataset-B. Our proposed model performs better than other state-of-the-art models regarding IoU, Dice score, and sensitivity with BUSI and Dataset-B. Though the number of parameters is large, the number of Floating-point operations per second (FLOPS) is lower, suggesting that our approach is simpler and requires no intricate design. Visual outputs of the proposed model and other state-of-the-art models are shown in Fig. 4. Our approach captures the better spatial structure of the breast lesions when compared to other state-of-the-art models.

Table 6. The External validation segmentation results for the STU dataset with baseline and proposed model trained with BUSI and Dataset-B. Sen, Prec, and Spec represent sensitivity, precision and specificity, respectively.

Models	BUSI					Dataset-B				
	IoU	Dice	Sen.	Prec.	Spec.	IoU	Dice	Sen.	Prec.	Spec.
UNet	69.77 ± 2.21	82.18 ± 1.54	79.11 ± 4.31	85.72 ± 2.82	98.16 ± 0.50	57.82 ± 6.78	73.09 ± 5.55	59.76 ± 7.54	94.92 ± 1.04	99.54 ± 0.14
Attention UNet	73.36 ± 3.69	84.60 ± 2.49	85.83 ± 0.65	83.48 ± 4.55	97.62 ± 0.79	68.43 ± 3.35	81.23 ± 2.32	72.71 ± 2.10	88.67 ± 3.48	99.12 ± 0.40
UNet ++	74.02 ± 2.33	80.05 ± 10.90	83.18 ± 2.47	87.04 ± 1.16	98.29 ± 0.17	68.11 ± 2.77	81.01 ± 1.95	73.83 ± 4.41	90.08 ± 3.95	98.80 ± 0.57
UNet 3+	71.63 ± 1.23	83.46 ± 0.84	81.65 ± 2.15	85.42 ± 1.62	98.07 ± 0.28	66.38 ± 1.90	79.78 ± 1.39	71.11 ± 1.49	92.25 ± 1.64	99.17 ± 0.21
SegNet	75.13 ± 0.62	85.80 ± 0.41	85.26 ± 2.22	86.45 ± 2.40	98.14 ± 0.44	68.30 ± 2.63	78.99 ± 1.92	69.30 ± 3.37	91.97 ± 1.90	99.15 ± 0.23
AAUNet	75.41 ± 3.22	85.96 ± 2.11	84.72 ± 2.52	87.29 ± 3.08	98.28 ± 0.47	70.06 ± 2.66	82.38 ± 1.86	74.35 ± 3.30	92.45 ± 0.59	99.16 ± 0.10
Proposed Method	79.08 ± 1.28	88.32 ± 0.80	87.43 ± 0.71	89.23 ± 1.12	98.59 ± 0.16	74.79 ± 0.39	85.58 ± 0.25	79.48 ± 1.94	92.78 ± 2.02	99.13 ± 0.28

4.5 External Validation

External validation is a critical step in assessing the generalizability and robustness of segmentation models. In our study, we used the STU dataset as an external validation set. This dataset, acquired by different imaging systems and from different geographical locations compared to BUSI and Dataset-B, serves as an essential benchmark for evaluating our proposed model's performance in real-world scenarios. Testing our model on this external dataset ensures its effective generalization to unseen data and different acquisition conditions. We trained models using BUSI and Dataset-B separately and tested them with the STU dataset to predict whether each pixel in the BUS image is normal or a tumour. Our proposed model demonstrates better generalizability to the unseen datasets than other models, with the results shown in Table 6.

5 Conclusions

Our study introduces a novel UNet-based model integrating GCN blocks in skip connections to facilitate breast lesion segmentation in BUS images. Our proposed model demonstrates superior performance compared to existing state-of-the-art methods in this domain. Through several ablation studies, we explain the significance of individual model components, providing insights into their contributions to segmentation accuracy. Moreover, we emphasize the pivotal role of image preprocessing in enhancing segmentation performance for BUS images. Our model showcases robustness across unseen datasets. Looking ahead, we aim to extend our model's capabilities beyond segmentation to encompass comprehensive tasks such as cancer detection, identification, and segmentation within a unified framework.

Acknowledgements. The authors would like to express their sincere gratitude to Prof. Hema A. Murthy, Emeritus Professor, Department of Computer Science and Engineering, IIT Madras, for her invaluable guidance and support throughout the completion of this research. The authors also extend their heartfelt thanks to IITM Pravartak Technologies Foundation, a Technology Innovation Hub of the Indian Institute of Technology, Madras, funded by the Department of Science and Technology, Government of India, under its National Mission on Interdisciplinary Cyber-Physical Systems, for supporting Anand Thyagachandran through a fellowship grant.

References

1. Jemal, A., Bray, F., Center, M.M., Ferlay, J., Ward, E., Forman, D.: Global cancer statistics. CA: Cancer J. Clin. **61**(2), 69–90 (2011)
2. Chan, V., Perlas, A.: Basics of ultrasound imaging. Atlas of ultrasound-guided procedures in interventional pain management, pp. 13–19 (2011)
3. Zhou, Z., et al.: Semi-automatic breast ultrasound image segmentation based on mean shift and graph cuts. Ultrason. Imaging **36**(4), 256–276 (2014)
4. Pons, G., Martí, J., Martí, R., Ganau, S., Noble, J.A.: Breast-lesion segmentation combining b-mode and elastography ultrasound. Ultrason. Imaging **38**(3), 209–224 (2016)
5. Xian, M., Zhang, Y., Cheng, H.D., Xu, F., Zhang, B., Ding, J.: Automatic breast ultrasound image segmentation: a survey. Pattern Recogn. **79**, 340–355 (2018)
6. Chen, G., Li, L., Dai, Y., Zhang, J., Yap, M.H.: AAU-net: an adaptive attention u-net for breast lesions segmentation in ultrasound images. IEEE Trans. Med. Imaging (2022)
7. Yan, Y., Liu, Y., Wu, Y., Zhang, H., Zhang, Y., Meng, L.: Accurate segmentation of breast tumors using ae u-net with HDC model in ultrasound images. Biomed. Signal Process. Control **72**, 103299 (2022)

8. 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
9. Huang, H., et al.: Unet 3+: a full-scale connected unet for medical image segmentation. In: 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), ICASSP 2020, pp. 1055–1059. IEEE (2020)
10. Huang, R., et al.: Boundary-rendering network for breast lesion segmentation in ultrasound images. Med. Image Anal. **80**, 102478 (2022)
11. Oktay, O., B., et al.: Attention u-net: learning where to look for the pancreas. arXiv preprint arXiv:1804.03999 (2018)
12. Zhang, X., et al.: Attention to region: region-based integration-and-recalibration networks for nuclear cataract classification using as-oct images. Med. Image Anal. **80**, 102499 (2022)
13. Zhou, Z., Siddiquee, M.M.R., Tajbakhsh, N., Liang, J.: Unet++: redesigning skip connections to exploit multiscale features in image segmentation. IEEE Trans. Med. Imaging **39**(6), 1856–1867 (2019)
14. Almajalid, R., Shan, J., Du, Y., Zhang, M.: Development of a deep-learning-based method for breast ultrasound image segmentation. In: 2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA), pp. 1103–1108. IEEE (2018)
15. Byra, M., et al.: Breast mass segmentation in ultrasound with selective kernel u-net convolutional neural network. Biomed. Signal Process. Control **61**, 102027 (2020)
16. Chen, G., Li, L., Zhang, J., Dai, Y.: Rethinking the unpretentious u-net for medical ultrasound image segmentation. Pattern Recogn. **142**, 109728 (2023)
17. Shareef, B., Xian, M., Vakanski, A.: Stan: small tumor-aware network for breast ultrasound image segmentation. In: 2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI), pp. 1–5. IEEE (2020)
18. Chen, G., Liu, Y., Dai, Y., Zhang, J., Cui, L., Yin, X.: Bagnet: bidirectional aware guidance network for malignant breast lesions segmentation. In: 2022 7th Asia-Pacific Conference on Intelligent Robot Systems (ACIRS), pp. 112–116. IEEE (2022)
19. Lee, H., Park, J., Hwang, J.Y.: Channel attention module with multiscale grid average pooling for breast cancer segmentation in an ultrasound image. IEEE Trans. Ultrason. Ferroelectr. Freq. Control **67**(7), 1344–1353 (2020)
20. Abraham, N., Khan, N.M.: A novel focal Tversky loss function with improved attention u-net for lesion segmentation. In: 2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019), pp. 683–687. IEEE (2019)
21. Qin, X., Zhang, Z., Huang, C., Gao, C., Dehghan, M., Jagersand, M.: Basnet: boundary-aware salient object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 7479–7489 (2019)
22. Wang, Y., et al.: Deeply-supervised networks with threshold loss for cancer detection in automated breast ultrasound. IEEE Trans. Med. Imaging **39**(4), 866–876 (2019)
23. Punn, N.S., Agarwal, S.: RCA-IUnet: a residual cross-spatial attention-guided inception u-net model for tumor segmentation in breast ultrasound imaging. Mach. Vis. Appl. **33**(2), 27 (2022)
24. Zhuang, Z., Li, N., Joseph Raj, A.N., Mahesh, V.G., Qiu, S.: An RDAU-net model for lesion segmentation in breast ultrasound images. PLoS ONE **14**(8), e0221535 (2019)

25. 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)
26. Cao, X., Chen, H., Li, Y., Peng, Y., Wang, S., Cheng, L.: Dilated densely connected u-net with uncertainty focus loss for 3D ABUS mass segmentation. Comput. Methods Programs Biomed. **209**, 106313 (2021)
27. Chen, G., Dai, Y., Zhang, J.: C-net: cascaded convolutional neural network with global guidance and refinement residuals for breast ultrasound images segmentation. Comput. Methods Programs Biomed. **225**, 107086 (2022)
28. Chen, G., Yin, J., Dai, Y., Zhang, J., Yin, X., Cui, L.: A novel convolutional neural network for kidney ultrasound images segmentation. Comput. Methods Programs Biomed. **218**, 106712 (2022)
29. Irfan, R., Almazroi, A.A., Rauf, H.T., Damaševičius, R., Nasr, E.A., Abdelgawad, A.E.: Dilated semantic segmentation for breast ultrasonic lesion detection using parallel feature fusion. Diagnostics **11**(7), 1212 (2021)
30. Hu, Y., et al.: Automatic tumor segmentation in breast ultrasound images using a dilated fully convolutional network combined with an active contour model. Med. Phys. **46**(1), 215–228 (2019)
31. Li, C., Wang, X., Liu, W., Latecki, L.J., Wang, B., Huang, J.: Weakly supervised mitosis detection in breast histopathology images using concentric loss. Med. Image Anal. **53**, 165–178 (2019)
32. 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)
33. Woo, S., Park, J., Lee, J.Y., Kweon, I.S.: CBAM: convolutional block attention module. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 3–19 (2018)
34. Xue, C., et al.: Global guidance network for breast lesion segmentation in ultrasound images. Med. Image Anal. **70**, 101989 (2021)
35. Peng, C., Zhang, X., Yu, G., Luo, G., Sun, J.: Large kernel matters–improve semantic segmentation by global convolutional network. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4353–4361 (2017)
36. Li, M.Y., Zhu, D.J., Xu, W., Lin, Y.J., Yung, K.L., Ip, A.W.: Application of u-net with global convolution network module in computer-aided tongue diagnosis. J. Healthcare Eng. **2021**(1), 5853128 (2021)
37. Ulyanov, D., Vedaldi, A., Lempitsky, V.: Improved texture networks: maximizing quality and diversity in feed-forward stylization and texture synthesis. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 6924–6932 (2017)
38. Hooley, R.J., Scoutt, L.M., Philpotts, L.E.: Breast ultrasonography: state of the art. Radiology **268**(3), 642–659 (2013)
39. Gonzalez, R.C.: Digital Image Processing. Pearson Education India (2009)
40. Zhao, Z., Yang, L., Long, S., Pi, J., Zhou, L., Wang, J.: Augmentation matters: a simple-yet-effective approach to semi-supervised semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 11350–11359 (2023)
41. Hofmanninger, J., Prayer, F., Pan, J., Röhrich, S., Prosch, H., Langs, G.: Automatic lung segmentation in routine imaging is primarily a data diversity problem, not a methodology problem. Eur. Radiol. Exp. **4**, 1–13 (2020)

42. De Boer, P.T., Kroese, D.P., Mannor, S., Rubinstein, R.Y.: A tutorial on the cross-entropy method. Ann. Oper. Res. **134**, 19–67 (2005)
43. Paszke, A., et al.: PyTorch: an imperative style, high-performance deep learning library. In: Advances in Neural Information Processing Systems, vol. 32 (2019)
44. Al-Dhabyani, W., Gomaa, M., Khaled, H., Fahmy, A.: Dataset of breast ultrasound images. Data Brief **28**, 104863 (2020)
45. Yap, M.H., et al.: Breast ultrasound region of interest detection and lesion localisation. Artif. Intell. Med. **107**, 101880 (2020)
46. Badrinarayanan, V., Kendall, A., Cipolla, R.: Segnet: a deep convolutional encoder-decoder architecture for image segmentation. IEEE Trans. Pattern Anal. Mach. Intell. **39**(12), 2481–2495 (2017)

SFRSeg-Net: Synovial Fluid Region Segmentation from Rheumatoid Arthritis Affected Small Joints Using USG for Early Detection

Puja Das[1], Sourav Dey Roy[1], Kaberi Sangma[1], Asim De[2], and Mrinal Kanti Bhowmik[1(✉)]

[1] Department of Computer Science and Engineering, Tripura University (A Central University), Suryamaninagar 799022, Tripura (W), India
mrinalkantibhowmik@tripurauniv.ac.in

[2] Department of Radiodiagnosis, Agartala Government Medical College, Govind Ballabh Pant Hospital, Government of Tripura, Agartala 799006, Tripura (W), India

Abstract. Synovial fluid imbalance in joints plays a significant role in the diagnosis of Rheumatoid arthritis (RA) at an early stage. RA mostly attacks the small joints like finger and wrist joints which makes it challenging to segment the synovial fluid from those small joints automatically. Although ultrasonography (USG) imaging is very sensitive to small joints and its fluid assessment, segmentation of synovial fluid regions from the USG images in the literature are less understood. Moreover, towards computer vision related research (especially segmentation of the suspicious abnormal regions) using USG imaging, several challenges exists including (a) USG images are prone to the certain artifacts in terms of noises because of which the presence of the different appearance of synovial fluid is less distinguishable with respect to the other anatomical appearances, (b) Also, with respect to imbalance occurrence, synovial fluid changes its shape, size and locations from one USG image to another USG image. To cope with such predefined challenges, we proposed a novel lightweight network named as "Synovial Fluid Region Segmentation Network (SFRSeg-Net)" for segmentation of synovial fluid regions from the USG images. The proposed network enriches by incorporating a novel Interpretable Element Wise Additive-Discrete Wavelet Transform (IEWA-DWT) based down sampling strategy to extract the significant salient features by ignoring noise imposed in USG imaging and maintain the original image integrity. As synovial fluid varies its shape in one image to another image, so boundary loss is also significant learning parameter with pixel wise region loss and its mutual combination is used as a loss function in our proposed network. Experimental results on publicly available USG imaging dataset reveal that our proposed SFRSeg-Net performed well with Dice similarity coefficient of 0.9066 $\pm$ 0.0520 which surpasses both the most recent state-of-the-art techniques and the current baseline.

Keywords: Rheumatoid Arthritis · Ultrasonography Images · Synovial Fluid · Segmentation Network · IEWA-DWT · Performance Evaluation

A. Antonacopoulos et al. (Eds.): ICPR 2024, LNCS 15313, pp. 127–146, 2025.
https://doi.org/10.1007/978-3-031-78201-5_9

1 Introduction

Approximately 0.92% of Indian adults suffer from rheumatoid arthritis [1]. The typical onset of rheumatoid arthritis (RA) is common for the age between of 35 to 60 years and it can also affect young children, even those under the age of sixteen [2]. According to the World Health Organization (WHO), RA affects women in about 70% of cases [3]. RA starts off affecting smaller joints such as fingers and toes and gradually spreads to the larger joints like elbows, knees, ankles, shoulders, hips and this disease spreads to the other organs too such as skin, eyes, heart, kidneys, and lungs [2, 4]. Consequently, RA is caused by a disease process that starts in the synovial membrane, the membrane that surrounds a joint and forms a protective hollow space [5]. This hollow space is filled with the liquid called synovial fluid [5]. The membrane swells up and produces extra synovial fluid, the swelling of the synovial membrane is called synovitis [6]. Synovial fluid's function is to lubricate the bone joint's cartilage and supply nutrition by diffusion so that there is less friction between the articular cartilages [6–9]. From this extensive study, it is observed that the disease significantly changes the synovial fluid at a severe stage as displayed in Fig. 1. Thus synovial fluid is the potential biomarker present in the affected joints which can help in decision making of disease diagnosis.

Fig. 1. Rheumatoid Arthritis affected hand joint digital camera visualization (a) and Ultrasound visualization (b)

To avoid the laborious manual work of the technicians and speed up the treatment process, the research community is deterministically working on building a computer aided diagnostic system for RA diagnosis. With the advancement of medical imaging equipments for computer aided disease diagnosis, the research community have utilized a variety of medical imaging modalities are used to diagnose RA, such as ultrasonography, computed tomography, Magnetic Resonance Imaging (MRI), Conventional Radiography, etc. [10]. When it comes to analysis of any fluid and small joints ultrasonography (USG) has been demonstrated to be more sensitive, real time and cost effective than clinical examination and other medical imaging modalities [29]. In addition, other markers of joint inflammation, such as synovitis, and marginal erosion, which may be radiologically occult, can be found and tracked with USG in the early stages of RA [11]. Although USG has the advancement in the literature of automatic detection of arthritis, there are still a lot of limitations using various image processing techniques. Among them noise is the most challenging one introduced into the ultrasound/ultrasonography images, which makes the presence of the suspicious region of interest (i.e., the appearance of synovial

fluid) appear less distinct than other anatomical appearances. In addition to these challenges, synovial fluid changes its shape and size from one ultrasound image to another ultrasound image. Therefore, segmentation of synovial fluid regions from the holistic USG images plays a crucial part in effective grading of arthritis.

As found in the literature [12–14], very limited work has been done on the segmentation of the USG images which are mostly based on conventional segmentation methods like active contour segmentation [12, 13]. Other than the conventional methods, in the last decades, Convolutional neural networks (CNNs) have been incredibly successful in the field of medical image analysis because of their ability to use detailed, important features with strong discriminating powers. Inspired by these insights, the paper aims at segmentation of the suspicious region of interest (ROI) from USG images thereby extracting the meaningful information about synovial fluid changes considering all the pre-defined limitations for RA diagnosis. In order to fulfill the specific necessities of the overall aim, we highlighted the contributions of this proposed paper below:

1. In this paper, we proposed a novel lightweight end-to-end deep neural network named as "Synovial Fluid Region Segmentation Network (SFRSeg-Net)" for segmenting the synovial fluid regions from USG imaging modality. As USG images are susceptible to various artifacts, specifically noise and acentric appearance of suspicious ROIs (i.e., synovial fluid regions), the network incorporates a novel Interpretable Element Wise Additive-Discrete Wavelet Transform (IEWA-DWT) based down sampling strategy so as to extract prominent features of the ROIs. This strategy increases the receptive field of lightweight networks, enabling them to capture contextual information of suspicious ROIs at long range. To the best of our knowledge, the proposed work is first attempt to use the building blocks of CNN for synovial fluid regions segmentation from RA affected small joints using USG images.
2. The experimental results on publicly available USG image dataset of RA patients [16] have been investigated and achieved new state-of-the-art results for extraction of synovial fluid regions from the holistic USG images.
3. Even though, the proposed SFRSeg-Net has shown consistent results in labeling the synovial and non-synovial regions precisely, the network also investigates the influences of various noises externally imposed in the USG images for synovial regions segmentation.

Paper Outline. Section 2 elaborates the review on the existing works for synovial fluid region segmentation from USG images. In Sect. 3, our proposed SFRSeg-Net is discussed in details. Section 4 reports the experimental results of the proposed network for synovial region segmentation. Finally, conclusion is provided in Sect. 5.

2 Related Work

In this section we discussed the existing literature related with our proposed task i.e., synovial fluid segmentation using Ultrasound imaging modality towards Rheumatoid arthritis detection. In the literature it has been observed that only three works have been done based on detection of synovial fluid towards rheumatoid arthritis diagnosis. In [12], Hemalatha et al. develop a method of segmentation of the synovial region based on active

contour technique using datasets from the MEDUSA database. The validation process reveals a rising true-positive rate, which averages 88.52% and ranges from 78.12% to 98.95% at its maximum. An average of 1.41% is removed from the false-positive rate. For instance, Veronese et al. [13] propose a method that can detect synovial borders in USG image semi-automatically with little to no user input using own datasets, a series of two distinct active contours is formed, the composition of which matches the entire synovial border. This method provides the sensitivity of 85 ± 13%, mean Dice's similarity index of 80 ± 8% and with a mean Hausdorff distance from the manual segmentation of 28 ± 10 pixels. In another paper of Hemalatha et al. [14], they have detected synovial fluid using various active contour segmentation methods on the database of MEDUSA. Performance analysis of their proposed approach shows Dice coefficient with 0.873 ± 0.005, Hausdorff distance with 18.7 ± 0.010. From existing literature of the similar task, it has been observed that most of the work focused on synovial region segmentation using traditional methods. However, the traditional methods adopted by the research community often suffers from the difficulties of parameter selection and accurate segmentation with a minimum rate of cumulative over and under segmentation. Moreover, the existing works discussed did not addresses the challenges of their used methods, algorithms and dataset. The existing tasks were implemented based on traditional and semiautomatic approaches. Therefore, there is huge scope to develop a deep learning network thereby overcoming the pre-defined challenges of USG images for synovial fluid segmentation towards RA diagnosis. To address this disparity, a novel segmentation network is proposed incorporating our proposed wavelet pooling based down sampling strategy for synovial fluid region segmentation from the USG images. In recent days, researchers has proposed few works focusing on incorporating wavelet based pooling in CNN frameworks for various vision based applications. In [23], Williams et al. introduced discrete wavelet pooling for convolutional neural networks to reduce features more structurally than other pooling via neighborhood regions while resolving the overfitting issue raised by max pooling. In this proposed discrete wavelet pooling mechanism, first-level high pass filter generated sub-bands are discarded and second-level sub-bands are considered to reduce feature dimensions further in the network. This method is validated on four benchmark datasets (MNIST, CIFAR-10, SHVN, and KDEF) for classification purposes. For instance, Souza et al. [46] used a combination of max-pooling and wavelet pooling (DWT) followed by 1×1 convolution for semantic segmentation by concatenating both outputs. This proposed method is used to solve the issue of loss of information caused by existing pooling methods which reduce the number of parameters, improve invariance to certain distortions, and enlarge the receptive field. IRRG images from the Potsdam and Vaihingen datasets have been used to validate this proposed wavelet-based pooling method. Consequently, to deal with noise interruption in convolutional neural networks (CNNs) for image classification, Li et al. [47] proposed to integrate CNNs with the simplest wavelet based pooling. This proposed method dropped the high-frequency components to reduce noise and improve image classification. This method was validated on the COCO dataset for suppression of the aliasing effect of noise. In the paper [48], to reduce noise in image segmentation, Zhao et al. substitute discrete wavelet transform for the conventional down-sampling modules. This proposed method splits the features into low- and high-frequency components and then removes the high-frequency components.

This method is validated on the Aneurysm dataset. From these reported works, we can infer that till date simplest discrete wavelet pooling is implemented in the field of different noise robust image segmentation and classification tasks. All of the above related work drops the high-frequency components where the problem of feature loss can arise. To address this issue, in our proposed segmentation network, we have designed a novel Interpretable Element Wise Additive-Discrete Wavelet Transform (IEWA-DWT) based down sampling strategy which considers some of the high frequency wavelet sub-bands thereby excluding the sub-band reflecting noise component.

3 SFRSeg-Net: Synovial Fluid Region Segmentation Network

In this section, we describe our proposed fully automated synovial fluid region segmentation network named as SFRSeg-Net which is based on the sequential applications of end-to-end encoder-decoder structure thereafter incorporating a novel interpretable element wise additive discrete wavelet transform (IEWA-DWT) based down sampling strategy so as to extract prominent features of the ROIs. Figure 2 illustrates the proposed segmentation network. Each of the components of the proposed network and its learning process is elaborated next.

3.1 Problem Definition

Let us define that input to our proposed network are USG images in RGB palette as I_t: $\varphi_t \rightarrow \mathbb{R}^3$, where t = 1, 2, …, t as the training images with size 128 × 128 × 3 and each of them having their corresponding ground truth defining the synovial fluid regions in the form of binary masks which is also defined as G_t: $\varphi_t \rightarrow \mathbb{R}^2$. We can state segmentation as the following optimization problem with respect to the network parameters γ and is mathematically expressed as Eq. (1):

$$\min_{r} \sum_{t} \sum_{i} L\big(G_t(i), O_{ft}(i, r)\big) \tag{1}$$

Here, $i \in \varphi_t$, $O_{ft}(i, \gamma)$ represents all the predicted output of the proposed optimized model by learning parameters γ. Our proposed model aims to learn all the parameters precisely towards correct pixel prediction by minimizing the loss between predicted pixel and ground truth pixel value and position.

3.2 Architecture Overview of the Proposed SFRSeg-Net

In this subsection, we discussed our proposed deep neural network for segmenting ROI i.e., synovial fluid regions from the holistic USG images towards RA detection. The proposed network named as "SFRSeg-Net (Synovial Fluid Region Segmentation Network)" consists of fully convolutional neural networks (FCNNs) with contracting and expanding paths which are popularly dubbed as down sampling and up sampling respectively. Finally, a softmax activation function was applied to the output of the final layer after the last up sampling block.

Module 1 (Down Sampling Path). For synovial fluid region segmentation, four blocks of layers in the down sampling path made up the proposed network as shown in Fig. 2. Each block designed with two convolution layers followed by batch normalization and the RELU activation function. After each block in the down sampling path as displayed in Fig. 2, we have incorporated the proposed IEWA-DWT as the pooling mechanism. Each of the layer constituting the down sampling path of the proposed networks are detailed below:

Convolutional Layer. The core building structure of FCNN is the convolution layer where multiple filters with tunable learnable weights and biases are used to extract low and high level distinguishable features to generate a feature map. Let $Z_n^{(m^{(l-1)})}$ be an input of l^{th} layer from the $(l-1)$ layer, k_{mn}^{l} be the learnable filter and b_m^l be the bias for the l^{th} layer. So for m^{th} output feature map, the n^{th} receptive filed from $(l-1)$ layer is convolved with m^{th} kernel of l^{th} layer and subsequently the bias added as shown in Eq. (2).

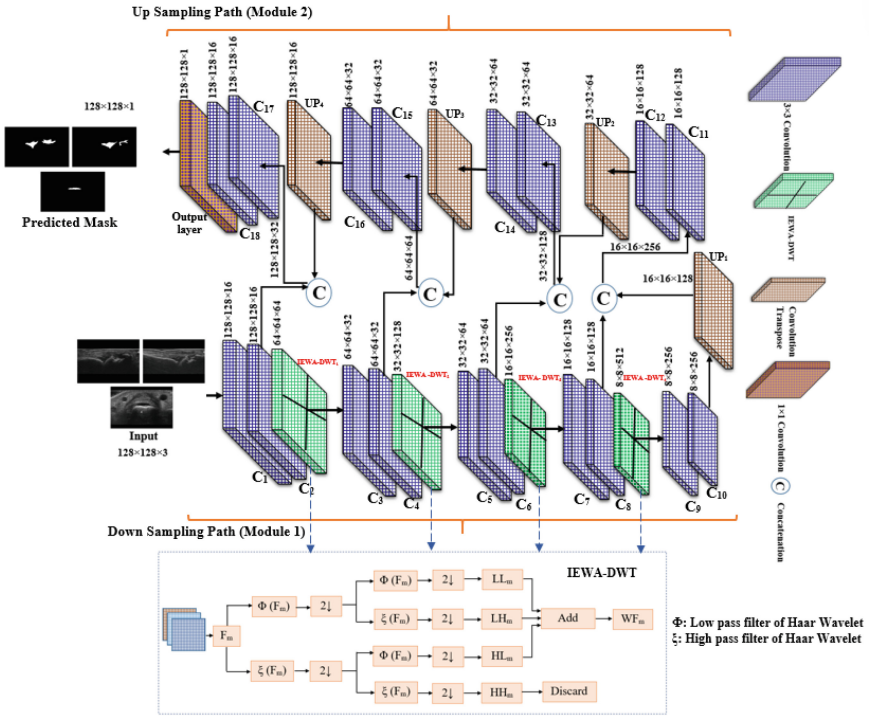


Fig. 2. Diagrammatic Representation of the Proposed SFRSeg-Net for Synovial Fluid Region Segmentation from the USG Images

$$C_p^l = \left(\sum_{n=1}^{j} Z_n^{m^{l-1}} * k_{mn}^l + b_m^l \right) \tag{2}$$

The out of this layer gone through an activation function which is non-linear in nature (in our proposed method it is Rectified Linear Unit (ReLu)) denoted as $\delta(\cdot)$. Our proposed down sampling strategy stacked up C_d convolution layers where $d \in \{1, 2, \ldots, 10\}$. Each of the C_d layers have k_p number of learnable kernels with size 3×3 where $p \in \{16, 32, 64, 128, 256\}$.

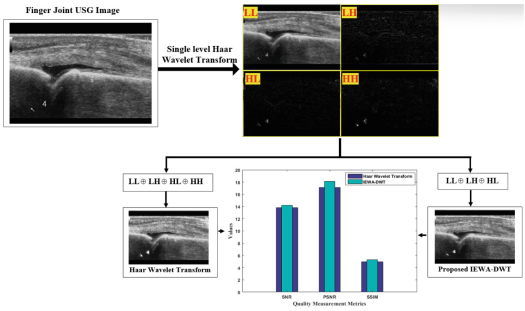


Fig. 3. Interpretation of Our Proposed IEWA-DWT Based Pooling for Synovial fluid Segmentation

Novel Interpretable Element Wise Additive Discrete Wavelet Transform (IEWA-DWT) Based Pooling. In order to decrease the dimensionality of the feature maps rep resenting the outputs of the intermediate activation layers and the number of subsequently learnable parameters, this layer individually conducts a down sampling operation over each activated feature map. Many existing networks addressed limitations of capability of learning long-range spatial dependencies with their down sampling strategies. In medical image segmentation, structural information is crucial and the existing pooling approaches do not preserve the structural information [30]. In addition, existing down sampling strategies are irreversible and invariably lead to information loss. Generally, synovial fluid changes its shape and size from image to image and therefore to solve this issue, our proposed network incorporates the novel idea of interpretable element wise additive discrete wavelet transform (IEWA-DWT) based pooling. The visual representation of the proposed IEWA-DWT as a pooling layer in our proposed network is illustrated in Fig. 2. In our proposed down sampling strategy, we have modify and design the wavelet block (IEWA-DWT) as a pooling layer in such a way that the network can learn long range spatial dependencies and also can ignore the noise of USG images in the progressive blocks and segment the region of interest with less computation. Our novel down sampling strategy potentializes on wavelet transforms which enable invertible down-sampling for self-attention learning and it ensuring the integrity of the data.

Inspired from [15], the proposed network reduces spatial information using IEWA-DWT while maintaining image directionality, which is independent of position, scale,

and rotation. This helps the proposed networks to generalize well and become resistant to overfitting by enabling the recognition features of fine-grained patterns or boundaries. Our proposed network consists of IEWA-DWT$_q$ number of pooling layers in each block of down sampling path where q ∈ {1, 2, …, 4} i.e., after 2nd, 4th, 6th, and 8th convolution layers. Considering the advantages of Haar wavelet [15] in our pooling layer, it will generate four sub-bands (LL, LH, HL and HH) of an input feature map with high pass (expressed in Eq. 3) and low pass (expressed in Eq. 4) filters. To mathematically express the strategy, let us assume that we have a feature map from certain convolution layer of the proposed network i.e., $F_m \in R^{H \times W \times C}$. Next, Discrete Wavelet Transform (DWT) has been employed using [15] to down sample the input feature maps in the pooling layers of our proposed network. In the DWT operation $\Pi(\bullet)$, firstly it will apply low pass (f_L) and high pass filter (f_H) in the row of the considered feature map to generate two bands and then same filters were employed on the produced two bands to further generates four sub bands (X_{LL}, X_{LH}, X_{HL}, X_{HH}) as shown in Eq. (5).

$$f_L = \left(\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}}\right) \tag{3}$$

$$f_H = \left(\frac{1}{\sqrt{2}}, -\frac{1}{\sqrt{2}}\right) \tag{4}$$

$$\Pi(F_m) = [X_{LL},\ X_{LH},\ X_{HL},\ X_{HH}] \tag{5}$$

In Eq. (5), X_{LL} denotes the low frequency component which reflects the structural information of the ROI from the holistic USG images under consideration. Consequently, the remaining X_{LH}, X_{HL}, X_{HH} are the high frequency components that reflect the texture and noise features respectively of the input feature maps. Thus, in our proposed neural network, we discarded the X_{HH} component to deal with the noise of the USG images. After that, we concatenate the three sub bands (X_{LL}, X_{LH}, X_{HL}) and this feature map transforms to successive blocks of layers in the down sampling path to extract local and global contextualized down-sampled feature maps. Let us consider F_m is the feature map generated from 2nd Convolution layer and it will progress through the first IEWA-DWT block which generates the transformed feature map (WF_m) with reduced dimension as shown in Eq. (6).

$$WF_m = X_{LL} \oplus X_{LH} \oplus X_{HL} \tag{6}$$

Here, $\oplus$ denotes the pixel wise addition of the generated sub bands. Figure 3 displays the diagrammatic interpretation of our proposed IEWA-DWT layer. To justify the effectiveness of our proposed layer (as displayed in Eq. (6)), we have analyzed the pixel wise additive approach of four bands with respect to our proposed approach. The analysis has been performed with respect to the assessment of the USG image quality as displayed in Fig. 3. The assessment has been performed with respect to three metrics including Signal to noise ratio (SNR), Peak signal to noise ratio (PSNR), and structural similarity index measure (SSIM). The range of SNR and PSNR start from 1 decibel (dB) and higher values indicate better image quality. Consequently, the range of SSIM is between −1 to 1 and value towards 1 indicates better image quality. For better visualization, the value

of the SSIM values has been plotted in 10 point scale thereby multiplying the metrics of the considered images with 10 for both approaches. It can be observed from Fig. 3 that our proposed IEWA-DWT (i.e., combined representation of X_{LL}, X_{LH}, and X_{HL} thereby discarding X_{HH}) have higher values of these considered metrics as compared to considering all the bands for pixel wise additive operation. This indicates that the corresponding intermediate feature representation using our proposed pooling layer (so as described in Fig. 2) can provide enhanced semantic representation of the USG images with respect to quality thereby reducing the noises.

Module 2 (Up Sampling Path). In the up sampling path of the proposed network, the output i.e., the significant content related to synovial fluid local features come and are up sampled by convolution transpose method. It consists of adding zeros to the input matrix to make it larger, followed by forward convolution based on the convolution kernel. Let consider after 10^{th} layer the feature map is $F_m^{aa,bb}$ and convolution kernel is K_m^a and the 8^{th} layer feature map is $F_m^{cc,dd}$ then to increase size of the $F_m^{aa,bb}$ zero will be added with (cc-aa, dd-bb) size. Then the obtained feature map i.e. $F_{mn}^{cc,dd}$ convolved with K_m^a. After that it will concatenate with the 8^{th} convolution layer output feature map i.e. $F_m^{cc,dd}$ to retrieve the location information of reconstructed significant pixels. In this way the up sample path is made by similar operation four times in the place after the 10^{th}, 12^{th}, 14^{th}, 16^{th} convolution layers respectively. Thereafter, two more coevolution layers were incorporated to reconstruct the image with its original size.

Output Layer. This layer in our proposed network performs the prediction of class labels. As our problem domain is based on binary class segmentation, we have designed our final layer with a sigmoid activation function [31] which considers real values ranging between 0 and 1. In the final layer i.e., after 18^{th} convolution layer of our proposed network, a sigmoid activation function is incorporated to generate the predicted binary maps representing the synovial fluid regions pertaining in the holistic USG images. The sigmoid activated output function can be mathematically expressed as Eq. (7).

$$O_f = Sigmoid\ (F_f) = \frac{1}{1 + e^{-F_f}} \tag{7}$$

Here, O_f is the output of the proposed model and F_f is the feature final map generated from the 18^{th} Convolution layer of the proposed network.

3.3 Training of Our Proposed SFRSeg-Net

In this subsection, we have illustrated the loss function which is utilized to optimize the proposed network by fine-tuning the learnable kernels. Depending upon the objective of the work (i.e., pixel based binary classification problem), we proposed to incorporate global loss function $L_{Global}(\bullet)$ in our network as a combination of binary cross entropy function $L_B(\bullet)$ and dice similarity loss function $L_D(\bullet)$ so as to learn the network parameters. Using a combination of both the loss allows the network to balance between pixel-wise accuracy and overall shape accuracy. Generally, the appearance of the regions of interest (i.e., synovial fluid regions in our proposed work) in ultrasound images differ significantly with blurred boundaries. Therefore, dice similarity loss function $L_D(\bullet)$

ensures that the predicted segmentation mask closely matches the ground truth in terms of shape and overlap. Consequently, to maximize the likelihood of the correct class (i.e., to indicate individual pixel misclassifications), we have used binary cross entropy $L_B(\bullet)$ as USG images suffer from low contrast so pixel wise classification is more necessary. Therefore, binary cross entropy loss function $L_B(\bullet)$ can predict the pixel wise similarity and dice similarity loss function $L_D(\bullet)$ ensures accurate boundary localization. The global loss function $L_{Global}(\bullet)$ of the proposed network is mathematically calculated as Eq. (8).

$$\begin{aligned} L_{Global}(G, O_f) &= L_B(G, O_f) + L_D(G, O_f) \\ &= \left[-\frac{1}{M} \sum_{i=1}^{M} G_i \cdot \log\left(O_{f_i}\right) + (1 - G_i) \log\left(1 - O_{f_i}\right) \right] + \left[\frac{1}{M} \sum_{i=1}^{M} \frac{2\left(G_i \times O_{f_i}\right)}{\left(G_i + O_{f_i}\right)} \right] \end{aligned} \tag{8}$$

Here, M represents the total number of samples (i.e., USG images) used for training the proposed network. The main goal of the proposed segmentation network is to minimize the loss function for every epoch thereby updating the learnable parameters as represented in Eq. (1). To continue the overall training process, ADAM optimizer [32] has been used with the tuned hyper parameters.

4 Results and Discussions

In this section, the results of the proposed network for segmentation of the synovial fluid regions from the holistic USG images are reported.

4.1 Dataset Details and Preparation

For measuring the effectiveness of the proposed network, we have used publicly available ultrasound images of RA subjects from [16]. The dataset comprises 49 USG images of the finger and wrist joints affected by RA which are available for the research community. Among these, 20 cases involve finger joints, with the remaining 29 cases concerning wrist joints. As our proposed work focuses on segmentation problems, therefore annotation of the suspicious ROIs defining the synovial fluid regions has been performed thereby reducing the strong subject biasness in annotating the ROIs. The ground truth annotation has been performed under the supervision of a medical experts using an open-source software tool called GNU Image Manipulation Programming, or GIMP [17]. Using the GIMP tool, the ROIs are annotated as white pixels representing the synovial fluid regions and black pixels representing the non-synovial regions of the USG images. Moreover, to reduce the strong subjective biases, maximum voting policy scheme is adopted similar to our previous work [33].

As our proposed network is a supervised learning approach, therefore to increasing the volume and variety of the dataset and thereafter generalizing the proposed SFRSeg-Net with respect to the USG image dataset, certain geometric based augmentation techniques are applied to the original dataset before serving to the proposed network. As

our dataset contains 49 USG images, 39 images are used for training and the remaining 10 images are used for testing the proposed model. Each of these training and testing sample subset of the dataset are independently performed augmentation to increase the size. For this, geometric augmentations including flipping (with horizontal and vertical flipping), rotation (with 30°, 45°, 75°, 90°, 115°, 135°, 150°, 175°, 200°, 225°, 250°, 270°, and 315°), rescaling (with 1/255), and zooming (with 0.3 and 0.5) are performed on the considered images and their corresponding ground truths. Depending upon these augmentation techniques, we have increased the training and testing data subsets as 741 and 180 augmented USG images respectively.

4.2 Implementation Details

As mentioned above, during training, Adam optimization [32] is utilized as the optimization algorithm. Experimentally, the network's learning rate, which determines how much the model's weights should change in response to the predicted error, is experimentally set at 0.001. Also, we experimentally set the momentum with a decaying learning rate to 0.9 and 0.005 respectively. For effective training, the training subset is divided in a 4:1 ratio to form a training set and validation set with their corresponding ground truth images. Total 741 images with their corresponding ground truth have been fed to the proposed SFRSeg-Net for the training process and 180 images are used to test the final optimized model. Before the training dataset (so as mentioned in Subsect. 4.1 of the manuscript) is sent into the network, random shuffling is performed. Also, for initiating the training process of the proposed network, the initialization of the kernel weights is done at random. Consequently, the input image size of the proposed network is 128 × 128 pixels in the time of training and testing. Also, we have trained our proposed network from 50 epoch to 300 varying epochs with a time stamp of 50 and mini-batch size of 16 and has been reported. To compare different state-of-the-art pooling mechanisms in the proposed SFRSeg-Net, our proposed IEWA-DWT layer is replaced by different pooling mechanisms in the respective positions of the proposed network and trained with similar hyper parameters as mentioned above. Using Tensorflow and Keras versions 2.10.0 [22], respectively, the proposed segmentation network was trained and tested entirely on a Python environment [21]. In addition, a 64 GB RAM in-stalled with NVIDIA GeForce GTX Titan XP GPU-based workstation (Model: HP Z4 Work-station) was utilized for the entire training and testing process.

4.3 Evaluation Metrics

In the existing literature, the segmentation method's performance are evaluated by Dice Similarity Coefficient (DSC) [18] and Jaccard Index (JI) [19] metrics. Therefore for fair comparison, our proposed work also evaluates using these metrics. The Dice Similarity Coefficient (DSC) and Jaccard index are statistical metrics that quantifies the similarity between two sets, often used in the context of image segmentation tasks. These metrics are widely used in the field of medical image analysis to assess the precision of automated segmentation algorithms by contrasting their outputs with a reference standard that is usually manually annotated by experts. Both DSC and JI produce a result between 0 and 1. A value of 0 says that there is no agreement between the two segmentations and no

overlap. Perfect overlap, or exactly matching the reference segmentation, is represented by a value of 1.

Table 1. Performance Analysis of our Proposed SFRSeg-Net with Varying Epochs

Network	Epoch(s)	DSC	JI
SFRSeg-Net [Our Proposed]	50	0.4095 ± 0.2127	0.2795 ± 0.1703
	100	**0.9066 ± 0.0520**	**0.8330 ± 0.0824**
	150	0.9022 ± 0.0550	0.8262 ± 0.0878
	200	0.9066 ± 0.0571	0.8338 ± 0.0901
	250	0.8971 ± 0.0582	0.8181 ± 0.0908
	300	0.9036 ± 0.0625	0.8296 ± 0.0985

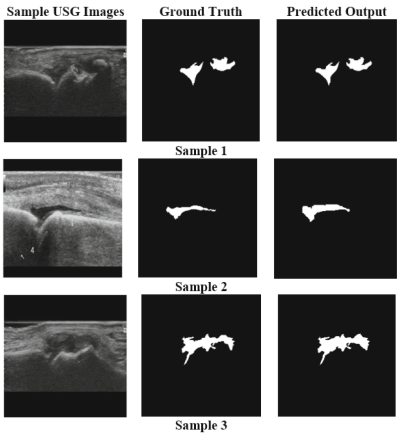


Fig. 4. Segmentation Results of the Proposed Model for Synovial Fluid Region Extraction from the USG Images

4.4 Segmentation Performance of the Proposed Network

In this subsection we have reported the performance of our proposed SFRSeg-Net on the testing subset of the used dataset so as described in Subsect. 4.1. The testing performance of the proposed network has been reported with the trained models obtained from varying epochs ranging from the 50 epochs to 300 epochs with time stamps of 50. All the hyperparameters remain similar for all the varied epochs so as described in Subsect. 4.2. The performance of the proposed model has been reported in Table 1 with varying epochs in terms of DSC and JI with standard representation (mean ± standard deviation). Based on these evaluation metrics, as reported in Table 1 which reveals that in 100 epoch the proposed model gives more precise segmentation with DSC of 0.9066 ± 0.0520 and

Table 2. Performance Comparison of the Proposed SFRSeg-Net with respect to State-of-the-Art Pooling Layer Mechanisms

Pooling Mechanism(s)		DSC	JI
Max Pooling	[24]	0.8082 ± 0.2061	0.6837 ± 0.2415
Average Pooling	[26]	0.5694 ± 0.2091	0.4280 ± 0.2103
Mixed Pooling	[25]	0.8958 ± 0.0591	0.8162 ± 0.0927
Polynomial Pooling	[28]	0.4975 ± 0.2116	0.4481 ± 0.2255
Second Order Pooling	[27]	0.8620 ± 0.2016	0.7574 ± 0.0580
Global Max Pooling	[34]	0.8855 ± 0.0582	0.7945 ± 0.1189
Stochastic Pooling	[35]	**0.8966 ± 0.1572**	**0.8126 ± 0.0782**
Global Average Pooling	[36]	0.8414 ± 0.1890	0.7262 ± 0.1569
IEWA-DWT (Our Proposed)		**0.9066 ± 0.0520**	**0.8330 ± 0.0824**

Jaccard Index (JI) of 0.8330 ± 0.0824. Figure 4 visualizes the qualitative results of the proposed network for synovial region segmentation for the best performed epoch.

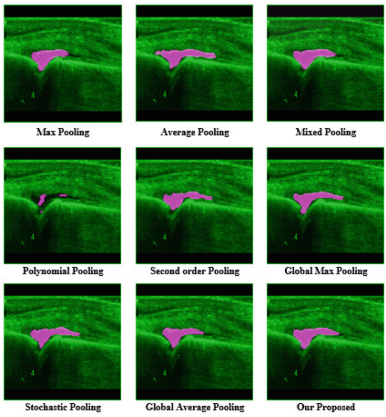


Fig. 5. Visualizing the Effectiveness of Segmented Synovial Fluid Regions from USG Finger Joint Image using Different Pooling Mechanisms in Our Proposed SFRSeg-Net.

4.5 Effectiveness of Different Pooling Mechanisms on Proposed Network

In this subsection, our proposed model is compared with eight different pooling mechanisms from the literature. These pooling mechanisms includes max pooling [24], mixed pooling [25], average pooling [26], second order pooling [27], polynomial pooling [28], global max pooling [34], stochastic pooling [35], and global average pooling [36]. Consequently, for fair comparison of our proposed network, we trained the proposed network

thereby replacing the proposed IEWA-DWT layer with the above mentioned eight pooling mechanisms. For fair comparison, the testing performance of the trained models with different pooling mechanisms for the best performed epoch has been reported in Table 2. For all the pooling mechanisms, except second order pooling [27], global max pooling [34], and stochastic pooling [35], the best training performance is obtained for 100 epochs. However, for second order pooling [27], global max pooling [34], and stochastic pooling [35] the best training performance has been observed to be for 150, 200, and 200 epochs. It can be observed from Table 2 that stochastic pooling mechanism [35] achieved the second-best result, resulting in an average DSC of 0.8966 ± 0.1572 and an average JI of 0.8126 ± 0.0782. Also, it has been observed that our proposed network incorporating the proposed IEWA-DWT demonstrated the most commendable performance at 100 epochs with an average DSC of 0.9066 ± 0.0520 and an average JI of 0.8330 ± 0.0824. In the Fig. 5, the results of the models (i.e. proposed SFRSeg-Net and replacing the proposed IEWA-DWT pooling layer with max pooling [24], mixed pooling [25], average pooling [26], second order pooling [27], polynomial pooling [28], global max pooling [34], stochastic pooling [35], and global average pooling [36] has been presented.

4.6 Comparison with the Existing State-of-the-Art Segmentation Methods

To verify the robustness of the proposed model, in this subsection, we compared the segmentation performance of the proposed SFRSeg-Net with respect to the state-of-the-art segmentation methods. The comparative study has been conducted in two major parts. In the first part, the state-of-the-art conventional and deep learning based segmentation methods are compared with our proposed network on the used datasets and is reported in Table 3. The methods used for comparison are active contour model (ACM) [37], Fuzzy C-means clustering (FCM) [38], K-means clustering [39], Level set method [40], Otsu thresholding [41], Watershed algorithm [42], SegNet [43], and U-Net [20]. The parameters of each of these compared methods are adjusted based on the recommendation of the authors and when not available are adjusted based on the enhanced results. In Table 3, the best performed segmentation method is represented as b old face and underline texts and consequently the second most best performed method is represented as bold face texts. It can be observed that among the state-of-the-art segmentation methods, U-Net has performed well for synovial fluid region extraction with an average DSC and JI of 0.8082 ± 0.2061 and 0.6837 ± 0.2415 respectively and can be considered as the second best performed method. Consequently, among all the compared method our proposed SFRSeg-Net has observed to obtain superior performance for segmentation of synovial fluid regions from the USG images with an average DSC and JI of 0.9066 ± 0.0520 and 0.8330 ± 0.0824 respectively.

In the second part, the existing methods used for the similar tasks (i.e., segmentation of imbalance of synovial fluid appearances from arthritis affected ultrasound images) are compared on the used dataset as reported in Table 4. Till date as mentioned in Sect. 2, very limited has addressed the task for synovial fluid region segmentation from the USG images towards RA detection. From literature, two competent methods as proposed by Veronese et al. [13] and Hemalatha et al. [14] are compared with our proposed SFRSeg-Net. The segmentation performance for each of these compared methods are reported

Table 3. Performance Comparison of Our Proposed Model with the State-of-the-Art Segmentation Methods on our Collected Publicly Available Dataset [16]

Segmentation Methods	DSC	JI
ACM [37]	0.2044 ± 0.0332	0.2567 ± 0.0491
FCM [38]	0.3906 ± 0.0661	0.3855 ± 0.0652
K-means [39]	0.3736 ± 0.0128	0.3525 ± 0.0483
Level set [40]	0.2392 ± 0.0497	0.2330 ± 0.0348
Otsu [41]	0.3554 ± 0.0485	0.3644 ± 0.0478
Watershed [42]	0.3673 ± 0.0716	0.3842 ± 0.0187
SegNet [43]	0.6964 ± 0.0537	0.5254 ± 0.0327
U-Net [20]	**0.8082 ± 0.2061**	**0.6837 ± 0.2415**
SFRSeg-Net (Our Proposed)	**0.9066 ± 0.0520**	**0.8330 ± 0.0824**

Table 4. Comparative analysis of our proposed SFRSeg-Net with the Existing State-of-the-Art Method for Synovial Fluid Region Segmentation

Methods	Dataset	DSC
Veronese et al. [13]	Own dataset	0.8000 ± 0.0080
	Public Dataset [16]	0.5889 ± 0.1036
Hemalatha et al. [14]	MEDUSA [44]	0.8730 ± 0.0050
	Public Dataset [16]	0.6458 ± 0.1357
SFRSeg-Net (Our Proposed)	Public Dataset [16]	**0.9066 ± 0.0520**

Table 5. Comparative Analysis of our Proposed SFRSeg-Net with the Existing Wavelet Pooling based Segmentation Methods

Compared Methods	Dataset	DSC
Souza et al. [46]	Public Dataset [16]	0.8530 ± 0.0050
Zhao et al. [48]		0.8708 ± 0.1092
SFRSeg-Net (Our Proposed)		**0.9066 ± 0.0520**

on the private datasets in terms of DSC. Therefore, for fair comparison, each of these considered methods are implemented and tested on the publically available USG image dataset [16] used in our proposed work. It can be observed that even though the compared methods has performed well for segmentation of the synovial fluid regions from the USG images, but our proposed SFRSeg-Net has shown superior seg mentation performance

on the used dataset with 0.32 and 0.26 percentage point improvements with respect to the methods proposed by Veronese et al. [13] and Hemalatha et al. [14] respectively.

Table 6. Performance of our Proposed SFRSeg-Net Model with respect to the Noise

Type of Noise	Method	DSC	JI
Gaussian	U-Net [20]	0.7492 ± 0.2969	0.6237 ± 0.2415
	SegNet [43]	0.6091 ± 0.3366	0.4938 ± 0.2783
	Our Proposed	**0. 8861 ± 0.2351**	**0.8228 ± 0.2053**
Poisson	U-Net [20]	0.7551 ± 0.3072	0.6459 ± 0.2498
	SegNet [43]	0.6254 ± 0.3300	0.5074 ± 0.2792
	Our Proposed	**0.8954 ± 0.2270**	**0.8394 ± 0.2792**
Salt & Paper	U-Net [20]	0.7322 ± 0.2970	0.6097 ± 0.2403
	SegNet [43]	0.5989 ± 0.3425	0.4810 ± 0.2840
	Our Proposed	**0.8809 ± 0.2229**	**0.8177 ± 0.3140**
Speckle	U-Net [20]	0.7387 ± 0.3125	0.6139 ± 0.2566
	SegNet [43]	0.6003 ± 0.3322	0.4875 ± 0.2811
	Our Proposed	**0.8818 ± 0.3240**	**0.8180 ± 0.2311**

To verify the robustness of the proposed segmentation network, the existing wavelet pooling based deep learning methods used for segmentation tasks as proposed by Souza et al. [46] and Zhao et al. [48] are also compared and reported in Table 5. For fair comparison, each of these methods are trained and tested on the same dataset [16] used for our proposed network. The training and testing datasets along with machine specifications remains similar as mentioned in Subsect. 4.2 used for implementing our proposed SFRSeg-Net. The hyper parameters of the compared methods were adjusted based on the recommendation of the respective authors and when not available are adjusted based on the enhanced results. From Table 5, it has been observed that the method so as proposed by Souza et al. [46] and and Zhao et al. [48] achieved DSC of 0.8530 ± 0.0050 and 0.8708 ± 0.1092 respectively on the similar testing set mentioned in Subsect. 4.2. Moreover, it can be seen from Table 5 that even though the compared methods has performed well for segmentation of the synovial fluid regions from the USG images, but our proposed SFRSeg-Net has shown superior segmentation performance on the used dataset with respect to the compared methods with DSC of 0.9066 ± 0.0520.

4.7 Impact of Noise in the Segmentation Performance of the Proposed SFRSeg-Net

Towards investigating the effectiveness of the segmentation performance of our proposed network in case of presence of noise in the USG images, different noises i.e., Gaussian Noise [45], Salt and Pepper Noise [45], Poisson Noise [45], and Speckle Noise [45] were

externally imposed into the test set of the ultrasound images. In our proposed work, the investigation was conducted with noise variance of 0.01 (i.e., Speckle and Gaussian noise). Table 6 reported the segmentation performance on noise imposed testing set of the USG images (so as mentioned in Subsect. 4.1) of the proposed SFRSeg-Net model. Also, for fair comparison, segmentation performance of the two best performed state-of-the-art segmentation methods i.e., U-Net [20] and SegNet [43] so as observed in Table 3 has been reported in Table 6. It is observed from Table 5, that there is decrease in the segmentation performance for the compared state-of-the-art methods and our proposed model. Moreover, it can also be observed from Table 6 that as compared to the competent methods [20, 43], the proposed model has performed better segmentation results for all the considered types of noises with an average DSC of 0.8860.

5 Conclusion

Towards diagnosis of RA, synovial fluids play a significant sign in the early stage. The paper proposed a novel SFRSeg-Net for automatic synovial fluid region segmentation from the USG images of small joints as small joints are very prone to RA and mostly early stage sign found in those joints only. We have observed that the proposed network provides better results concerning all the state-of-the-art methods including traditional and deep learning based methods. In this work, the proposed network also shows its capability to tolerate different noise oriented issues in segmentation tasks. This task can lead to building a computer aided decision making system by extracting the features from the segmented region and comparing it with the healthy subjects. Even though experimental results reveal that our proposed SFRSeg-Net outperforms for synovial regions segmentation from USG image, there are certain limitations that need to be addressed in future. As convolutional operations in our proposed network are limited to global con-text, so some new modules such as transformer based architectures can to be introduced which can deal with the local context of the USG images. On the other hand, towards reducing the noise from the USG images, SFRSeg-Net removes one high frequency sub band which may consist of some significant edge feature of synovial fluid regions. So these limitations may be resolved in future with some attention based mechanisms. In the future, the proposed work will be extended thereafter morphological features based analysis of the segmented ROIs toward severity prediction.

Acknowledgement. The work presented herein was being conducted at the Bio-Medical Infrared Image Processing Laboratory of the Computer Science and Engineering Department, Tripura University (A Central University), Tripura (W), India. This work was supported by the Department of Biotechnology (DBT), Government of India, under Grant No. BT/PR33087/BID/7/889/2019, Dated: 23/03/2022. The first author is grateful to Department of Science and Technology (DST), Government of India for providing DST INSPIRE Fellowship under Grant No. IF200476. The corresponding author, Mrinal Kanti Bhowmik is grateful to Short Term ICMR-DHR International Fellowship Programme for the partial support (Grant No. INDO/FRC/452/S69/2019-20-IHD, Dated: 09/12/2019).

References

1. Rheumatoid Arthritis. https://www.arthritis-india.com/rheumatoid-arthritis.html. Accessed 16 Jan 2024
2. Bullock, J., et al.: Rheumatoid arthritis: a brief overview of the treatment. Med. Princ. Pract. **27**(6), 501–507 (2019)
3. Rheumatoid arthritis. https://www.who.int/news-room/fact-sheets/detail/rheumatoid-arthritis. Accessed 20 Jan 2024
4. Rheumatoid arthritis. https://www.mayoclinic.org/diseases-conditions/rheumatoid-arthritis/symptoms-causes/syc-20353648. Accessed 28 Jan 2024
5. Sailaja, A.K.: An overall review on rheumatoid arthritis. J. Curr. Pharma Res. **4**(2), 1138 (2014)
6. Too Much Synovial Fluid: A Cause for Rheumatoid Arthritis. https://pcwfl.com/too-much-synovial-fluid-a-cause-for-rheumatoid-arthritis-pain-what-is-rheumatoid-arthritis/. Accessed 29 Jan 2024
7. Tamer, T.M.: Hyaluronan and synovial joint: function, distribution and healing. Interdiscip. Toxicol. **6**(3), 111–125 (2013)
8. Vincent, H.K., Percival, S.S., Conrad, B.P., Seay, A.N., Montero, C., Vincent, K.R.: Hyaluronic acid (HA) viscosupplementation on synovial fluid inflammation in knee osteoarthritis: a pilot study. Open Orthop. J. **7**, 378 (2013)
9. Vernasca, C., Giori, A.M., Togni, S.: U.S. Patent Application No. 13/976,113 (2014)
10. Mota, L.M.H.D., et al.: Imaging diagnosis of early rheumatoid arthritis. Rev. Bras. Reumatol. **52**, 761–766 (2012)
11. Saran, S., Bagarhatta, M., Saigal, R.: Diagnostic accuracy of ultrasonography in detection of destructive changes in small joints of hands in patients of rheumatoid arthritis: a comparison with magnetic resonance imaging. J. Assoc. Physicians India **64**(11), 26–30 (2016)
12. Hemalatha, R.J., Vijaybaskar, V., Thamizhvani, T.R.: Automatic localization of anatomical regions in medical ultrasound images of rheumatoid arthritis using deep learning. Proc. Inst. Mech. Eng. Part H: J. Eng. Med. **233**(6), 657–667 (2019)
13. Veronese, E., et al.: Improved detection of synovial boundaries in ultrasound examination by using a cascade of active-contours. Med. Eng. Phys. **35**(2), 188–194 (2013)
14. Hemalatha, R.J., Vijaybaskar, V., Thamizhvani, T.R.: Performance evaluation of contour based segmentation methods for ultrasound images. Adv. Multimed. **2018**, 1–8 (2018). https://doi.org/10.1155/2018/4976372
15. Porwik, P., Lisowska, A.: The Haar-wavelet transform in digital image processing: its status and achievements. Mach. Graph. Vis. **13**(1/2), 79–98 (2004)
16. Ultrasound Cases. https://www.ultrasoundcases.info/. Accessed 22 Jan 2024
17. GIMP. https://www.gimp.org/. Accessed 27 Jan 2024
18. Bertels, J., et al.: Optimizing the dice score and Jaccard index for medical image segmentation: theory and practice. In: Shen, D., et al. (eds.) MICCAI 2019. LNCS, vol. 11765, pp. 92–100. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-32245-8_11
19. Costa, L.D.F.: Further generalizations of the Jaccard index. arXiv preprint arXiv:2110.09619 (2021)
20. Ronneberger, O., Fischer, P., Brox, T.: U-net: convolutional networks for biomedical image segmentation. In: Navab, N., Hornegger, J., Wells, W., Frangi, A. (eds.) MICCAI 2015. LNCS, vol. 9351, pp. 234–241. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-24574-4_28
21. Python 3.9.7. https://www.python.org/downloads/release/python-397/. Accessed 02 Feb 2024
22. Install TensorFlow 2. https://www.tensorflow.org/install. Accessed 03 Feb 2024

23. Williams, T., Li, R.: Wavelet pooling for convolutional neural networks. In: International Conference on Learning Representations (2018)
24. Nagi, J., et al.: Max-pooling convolutional neural networks for vision-based hand gesture recognition. In: 2011 IEEE International Conference on Signal and Image Processing Applications (ICSIPA), pp. 342–347. IEEE (2011)
25. Yu, D., Wang, H., Chen, P., Wei, Z.: Mixed pooling for convolutional neural networks. In: Miao, D., Pedrycz, W., Ślęzak, D., Peters, G., Hu, Q., Wang, R. (eds.) RSKT 2014. LNCS, vol. 8818, pp. 364–375. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-11740-9_34
26. Yang, J., Xie, F., Fan, H., Jiang, Z., Liu, J.: Classification for dermoscopy images using convolutional neural networks based on region average pooling. IEEE Access **6**, 65130–65138 (2018)
27. Carreira, J., Caseiro, R., Batista, J., Sminchisescu, C.: Semantic segmentation with second-order pooling. In: Fitzgibbon, A., Lazebnik, S., Perona, P., Sato, Y., Schmid, C. (eds.) ECCV 2012. LNCS, vol. 7578, pp. 430–443. Springer, Heidelberg (2012). https://doi.org/10.1007/978-3-642-33786-4_32
28. Wei, Z., et al.: Building detail-sensitive semantic segmentation networks with polynomial pooling. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 7115–7123 (2019)
29. Perić, P., Pervan, M.: Diagnostic ultrasound of the small joints of the hands and feet: current status and role of ultrasound in early arthritis. Reumatizam **57**(2), 68–78 (2010)
30. Ghafoorian, M., et al.: Location sensitive deep convolutional neural networks for segmentation of white matter hyperintensities. Sci. Rep. **7**(1), 5110 (2017)
31. Narayan, S.: The generalized sigmoid activation function: competitive supervised learning. Inf. Sci. **99**(1–2), 69–82 (1997)
32. Kingma, D.P., Ba, J. Adam: a method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
33. Das, P., Marak, A., Roy, S.D., Gupta, R., Bhowmik, M.K.: Inflammatory bone region segmentation using USG Rheumatoid Arthritic images. In: 2023 14th International Conference on Computing Communication and Networking Technologies (ICCCNT), pp. 1–7. IEEE (2023)
34. Kim, H., Jeong, Y.S.: Sentiment classification using convolutional neural networks. Appl. Sci. **9**(11), 2347 (2019)
35. Zeiler, M.D., Fergus, R.: Stochastic pooling for regularization of deep convolutional neural networks. arXiv preprint arXiv:1301.3557 (2013)
36. Lin, M., Chen, Q., Yan, S.: Network in network. arXiv preprint arXiv:1312.4400 (2013)
37. Menet, S., Saint-Marc, P., Medioni, G.: Active contour models: overview, implementation and applications. In: 1990 IEEE International Conference on Systems, Man, and Cybernetics Conference Proceedings, pp. 194–199. IEEE (1990)
38. Peizhuang, W.: Pattern recognition with fuzzy objective function algorithms (James C. Bezdek). SIAM Rev. **25**(3), 442–442 (1983). https://doi.org/10.1137/1025116
39. Mucha, H.J., Späth, H.: Cluster dissection and analysis: theory, FORTRAN programs, examples. (Translator: Johannes Goldschmidt.) Ellis Horwood Ltd. Wiley, Chichester 1985, p. 226 (1986)
40. Gray, A., Abbena, E., Salamon, S.: Modern differential geometry of curves and surfaces with Mathematica (2006)
41. Otsu, N.: A threshold selection method from gray-level histograms. Automatica **11**(285–296), 23–27 (1975)
42. Meyer, F.: Topographic distance and watershed lines. Signal Process. **38**(1), 113–125 (1994)
43. Badrinarayanan, V., Kendall, A., Cipolla, R.: Segnet: a deep convolutional encoder-decoder architecture for image segmentation. IEEE Trans. Pattern Anal. Mach. Intell. **39**(12), 2481–2495 (2017)

44. About the project. https://medusa.aei.polsl.pl/. Accessed 26 Feb 2024
45. Rosin, P., Collomosse, J. (eds.): Image and Video-Based Artistic Stylisation. Springer, London (2013)
46. de Souza Brito, A., Vieira, M.B., De Andrade, M.L.S.C., Feitosa, R.Q., Gi-raldi, G.A.: Combining max-pooling and wavelet pooling strategies for semantic image segmentation. Expert Syst. Appl. **183**, 115403 (2021)
47. Li, Q., Shen, L., Guo, S., Lai, Z.: WaveCNet: wavelet integrated CNNs to suppress aliasing effect for noise-robust image classification. IEEE Trans. Image Process. **30**, 7074–7089 (2021)
48. Zhao, Y., Wang, S., Zhang, Y., Qiao, S., Zhang, M.: WRANet: wavelet integrated residual attention U-Net network for medical image segmentation. Complex Intell. Syst. **9**(6), 6971–6983 (2023)

Classification of Cutaneous Diseases: A Systematic Study on Real-Time Captured Images Using Deep Learning

Bhavik Kanekar[1], Jay Sawant[2], Niti Chikhale[2], Paras Dhotre[2], Sushil Savant[3], Gajanan Nagare[2], and Kshitij Jadhav[1(✉)]

[1] Koita Centre for Digital Health, Indian Institute of Technology, Bombay, Mumbai, India
{bhavikkanekar,kshitij.jadhav}@iitb.ac.in
[2] Department of Biomedical Engineering, Vidyalankar Institute of Technology, Mumbai, India
[3] The Humanitarian Clinic, Mumbai, India

Abstract. Dermatological infectious diseases pose a significant public health concern due to their highly contagious nature, often characterized by painful sores, fluid-filled blisters, flesh-colored bumps, and itching. Despite their distinct visual symptoms, diagnosing these diseases can be challenging due to their phenotypical similarities and overlapping clinical presentations with other skin conditions. In this paper we introduce a novel and diverse skin lesion dataset comprising patients from India, focusing on prevalent infectious skin conditions such as herpes zoster, herpes simplex, molluscum contagiosum, and non-viral skin disorders. These conditions are particularly common in this geographical region. Furthermore, we test Deep Learning models using different augmentation techniques, analyze the performance, and evaluate several metrics on different deep models using image augmentations. By assessing the performance of these models and analyzing several metrics across different augmentation methods, our findings demonstrate the capability of deep-learning models in classifying skin images and such computational techniques can be used to enhance healthcare accessibility and effectiveness, in resource-constrained settings like India.

Keywords: Cutaneous diseases · Deep learning models · Generative models

1 Introduction

Skin is part of the integumentary system and is the largest outer organ of the body providing us with sensory information and protecting internal organs from abnormal temperature and stress working as a protective shield to the body [23]. Skin diseases or cutaneous diseases are one of the most common causes of human illness [19] and demonstrate a wide array of conditions ranging from lesions, fluid-filled blisters, scaly patches, inflammation, flesh-colored bumps, mild irritation to chronic disfigurement. Due to these dermatological disfiguring conditions, patients suffer from mental health issues due to stigma leading to social exclusion or isolation [9]. These skin diseases can be caused due to genetic reasons in addition to infectious agents such as bacteria, fungi,

A. Antonacopoulos et al. (Eds.): ICPR 2024, LNCS 15313, pp. 147–162, 2025.
https://doi.org/10.1007/978-3-031-78201-5_10

viruses, and yeasts. The screening test for these cutaneous conditions is visual observation where a dermatologist inspects the affected area looking for specific characteristics such as texture, color and lesion distribution [29]. However, due to overlapping characteristics, phenotypical similarity and similar clinical presentations it is very difficult to reach an accurate differential diagnosis without sufficient experience which could lead to misdiagnosis [26].

In recent times, the advancement of deep learning models has shown tremendous success in the medical domain. Due to the availability of a large amount of annotated data and computing power, the performance of these deep models has shown substantial improvement. This is possible due to the large datasets available due to public and private data collections. There are numerous skin image datasets available online covering a wide variety of skin diseases. The open source dataset published in [7,17,35] and [12] provides a large number of benign and cancerous categories of skin pigmentation. Various state-of-the-art deep learning models already show remarkable classification performance on the above-mentioned datasets [18]. However, given the different skin texture and color, models developed on the Caucasian population would prove to be ineffective on the Indian population prompting the requirement of developing models on Indian datasets.

2 Our Contribution

The main contributions of the work are listed below.

- Literature review has demonstrated that currently there is no dataset available specific to Herpes, Molluscum contagiosum or Non-viral disease. We create and publish a skin lesion dataset for the Herpes simplex, Herpes zoster, Molluscum contagiosum and Non-viral, particularly on the Indian demographic. The dataset consists of 200 skin lesion images with 50 images in each category.
- We investigate various augmentation techniques, from simple operations like flipping and rotating, to more compute-intensive methods such as Wasserstein GAN (wGAN) and convolutional autoencoders. The results show that simple augmentation techniques give better results than model-based augmentation techniques.

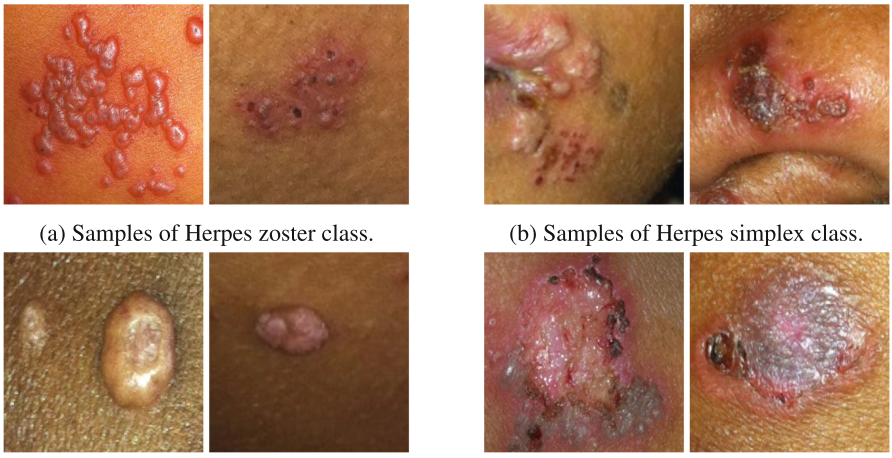
The structure of the paper is as follows: In Sect. 3 we discuss the previous work related to applications of machine learning for skin disease classification/ diagnosis. In Sect. 4 we provide the theoretical background for image augmentation techniques, convolutional neural network (CNN), wGAN and autoencoders. Further, in Sect. 5 we discuss the dataset while in Sect. 6 we elaborate on the methodology and discuss the results and conclude the study in Sect. 7.

3 Related Work

The study performed by [28] uses features extracted from images to classify skin disease. The Haar feature [33], color features and gray level co-occurrence matrix features were used. This study compares support vector machine (SVM), K-nearest neighbor

Table 1. Samples of the original images taken and extracted image lesions.

Class	Original Image	Extracted Image
Herpes Zoster		
Herpes Simplex		
Molluscum Contagiosum		
Non-Viral		



(a) Samples of Herpes zoster class.

(b) Samples of Herpes simplex class.

(c) Samples of molluscum contagiosum class.

(d) Samples of Non-viral class.

Fig. 1. Sample images from each class of our dataset.

and K-means clustering for the classification. Due to the availability of huge labeled dataset in this field, deep learning models have better results than traditional machine learning algorithms. In [21] multiple deep convolutional neural network (CNN) models for classifying four common fungal skin diseases were studied. The images underwent normalization and basic augmentation techniques namely cropping, rotations, flipping, translations, contrast adjustment, and scaling. The highest accuracy achieved is 93.3%. The work done in [19] proposes a smartphone-based application used for fungal disease classification using images. The MobileNet-V2 [24] CNN model is used for this purpose and the dataset used for this purpose is indigenous to the study location. The study performed by [29] uses spectral centroid magnitude for a feature extraction approach to classify skin diseases. This approach yielded a classification accuracy of 97% which exceeded other studies. Further, [31] proposed a method applying autoencoder, MobileNetV2, and spiking neural networks (SNN) for skin cancer detection. The experiments were conducted using the ISIC skin cancer dataset. The proposed method used features extracted using MobileNet-V2 on the original dataset and reconstructed dataset using autoencoder. These features were then combined and passed to the SNN model for classification. The study achieved a classification success rate of 95.27%. The authors in [10] propose a decision system model for medical experts for the diagnosis of skin lesion images. This novel method combines the texture features computed from images and visual attributes provided by the physician to make individual predictions and provide decisions on the majority vote for predictions.

After the development of GANs [11], wGANs [3] and variational autoencoders [5] these generative models are used for data augmentation and to handle class imbalance problem. In [1] authors used the CGAN techniques to solve the class imbalance issue by generating the desired images. With suitable data augmentation techniques, the suggested models achieved accuracies of 92% for VGG16, 92% accuracy using the ResNet-

50, and 92.25% accuracy using ResNet-101. For this study, the HAM10000 dataset is used. The study performed by [14] implemented a Deep Generative Adversarial Network (DGAN) for the classification of skin disorders. To address the class imbalance problem, different images from various images have been taken from datasets available online. To evaluate the effectiveness of GANs, two CNN models were simultaneously developed using the ResNet50 and VGG16 architectures. The training datasets were augmented using conventional rotation, flipping, and scaling techniques. DGAN surpassed conventional data augmentation methods, achieving a performance of 91.1% for the unlabelled dataset and 92.3% for the labeled dataset. Authors in [4] proposed a system for creating a robust skin condition identification system for Herpes Zoster diagnosis by condensing a group of Deep Neural Networks (DNN). This approach evaluated robustness using the proposed knowledge distillation from the ensemble via the curriculum training (KDE-CT) method. The main idea of this method was to train the teacher model on an ensemble of multiple DNNs. The curriculum training is used such that the student model can learn from a stronger teacher model. The Results concluded that the trained MobileNetV3-Small achieved better results (93.5% overall accuracy, 67.6 mean error) than the DNN ensemble.

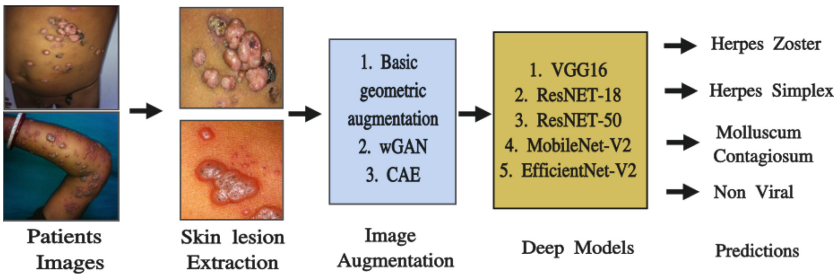


Fig. 2. The flowchart of the methodology.

4 Preliminary

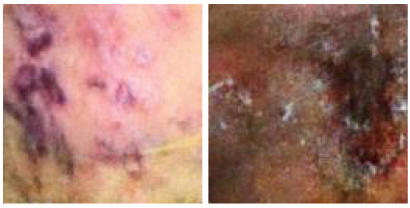
4.1 Basic Image Augmentation Techniques

The basic image augmentation techniques consist of geometrical transformations of images such as rotation, flipping, cropping etc. Even though these techniques are easy to implement the basic assumption is the distribution of the existing data is similar to real-world data [34].

4.2 Convolutional Neural Network

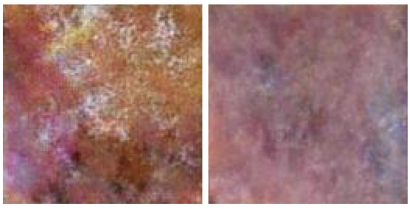
A CNN is a widely implemented computer vision algorithm for feature representation. CNN consists of convolution layers and pooling layers. In this case, the convolution layer uses multiple filters to extract the features from the input map, a multi-channel image. Let x_m^i represents the m-th input map at the layer i, the n-th output map y_n^i

(a) Samples of Herpes zoster class.



(b) Samples of Herpes simplex class.

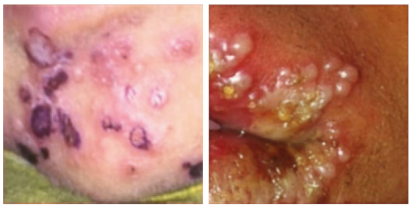
(c) Samples of molluscum contagiosum class.



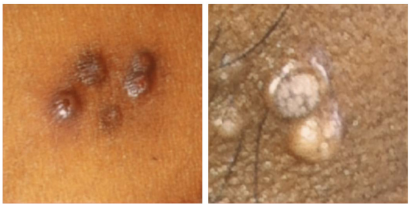
(d) Samples of Non-viral class.

Fig. 3. Sample images generated by WGAN for each class.

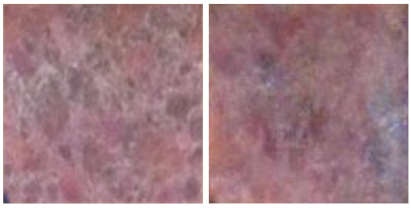
(a) Samples of Herpes zoster class.



(b) Samples of Herpes simplex class.



(c) Samples of Molluscum contagiosum class.



(d) Samples of Non-viral class.

Fig. 4. Sample images generated by CAE for each class.

at the i-th layer is represented in Eq. 1. The activation function rectified linear unit (ReLU).

$$y_i^n = ReLU\left(\sum_m^{M^{i-1}} w_i^{n,m} * x_m^i + b_n^i\right) \tag{1}$$

The pooling layer reduces the response provided to it to make it more compact. There are three types of pooling mix pooling, min pooling and average pooling.

4.3 wGAN

The deep learning-based GAN consists of a generator model G and discriminator model D. The data distribution is captured by 'G'. The discriminator 'D' gives the probability of the sample if it is generated by 'G' or provided from the training dataset. Both models are trained via an adversarial process simultaneously. The task of generator 'G' is to maximize the chances of 'D' to make an incorrect decision [20]. However, it is observed that the training of the GANs is very unstable since the updating of the generator becomes worse as the training progresses [6]. So the discriminator outperforms the generator, making the generator learn nothing due to mode collapse. wGAN tackles issues such as training instability and the vanishing gradient problem encountered in classical GANs [2]. This is because wGAN utilizes earth mover distance to compare the real and generated distribution [3]. Earth-Movers distance also known as the Wasserstein-1 metric is expressed in the Eq. 2.

$$W(p_g, p_r) = \inf_{y \in \prod(p_g, p_r)} E_{(x,x' \sim y)} ||x - x'|| \tag{2}$$

where $\prod(p_g, p_r)$ is the joint distributions $y(x, x')$ whose marginals are p_g and p_r.

4.4 Convolutional Autoencoder

An autoencoder consists of an encoder and a decoder. The encoder tries to convert the input data into feature space using a mapping function and non linearity [16]. The general function of autoencoder can be represented as Eq. 3, where x is input, b is bias, W is the weight of mapping function f and y is the hidden representation.

$$y = f(Wx + b). \tag{3}$$

Similarly, the decoder tries to predict the function f' to get z as the output also called as the reconstruction of x. This process is represented using 4. Here y is input, b' is bias, W is the weight of mapping function f' and z is the reconstruction.

$$z = f'(Wy + b'). \tag{4}$$

The main aim of the autoencoder is to reduce the reconstruction error achieved by reducing the cost function C_{ae}. The cost function C_{ae} is given by Eq. 5. Here p is the number of images, x_i-th and z_i-th are the corresponding input image and reconstructed image. $L[x_i, z_i]$ is the reconstruction loss that is expressed in the Eq. 6.

$$C_{ae} = \frac{1}{p} \sum_{i=1}^{p} L[x_i, z_i]. \tag{5}$$

$$L_{ae}[x_i, z_i] = ||x_i - z_i||^2. \tag{6}$$

Convolutional autoencoder (CAE) combines the convolution layer to autoencoder instead of connected layer [36]. Subsequently, the convolutional autoencoder performs the process of creating feature space from the input image and the convolutional decoder performs the reconstruction process of converting the feature space to the output. So the Eq. 3 and 4 can written as Eq. 7 and 8. Here w is the convolutional kernel for the encoder, x is the input, b is bias and y is the feature space.

$$y = ReLU(wx + b). \tag{7}$$

$$z = ReLU(w'y + b'). \tag{8}$$

where w' is the convolutional kernel for the decoder, y is the input and z is the reconstruction output.

4.5 Fréchet Inception Distance

Fréchet Inception Distance (FID) is a similarity measure between images [25]. It is shown that the measure correlates well with the visual quality of human judgment and has used InceptionNET to obtain visual-related features [15]. FID works using feature space representation rather than directly comparing pixel values. It first calculates statistics (mean and covariance) of feature representations (often called embeddings) extracted from real and generated/transformed images. The Fréchet distance is a similarity measure between two probability distributions in a metric space. In the context of FID, it's calculated between the multivariate Gaussian distributions formed by the statistics (mean and covariance) of feature representations of real and generated/transformed images. The Gaussian distribution can be referred to as the maximum entropy distribution for a given mean and covariance. The Fréchet Inception Distance is defined as the Fréchet Distance between the Gaussian distribution with mean (m, C) of original images and Gaussian with mean (m_w, C_w) of generated/transformed images. A lower FID indicates that the generated images are closer in terms of distribution to the real images in the feature space. Therefore, lower FID scores indicate better image quality. The FID equation consists of two parts, first part calculates the Euclidean distance between means of the feature representations of real and generated/transformed images. It measures how far apart are the average characteristics of the two sets of images. The second part calculates the trace of the covariance matrices of the real and generated/transformed images, adjusted for their covariance. It calculates the variance in the images' features, taking into account their interrelationships. The FID is given by Eq. 9, where m and m_w are the mean vector of feature representation of real images and generated images, C and C_w is the covariance of feature representation of real images and generated/transformed images and Tr denotes the sum of the diagonal elements of a matrix.

$$d^2((m, C), (m_w, C_w)) = ||m - m_w||_2^2 + Tr(C + C_w - 2(CC_w)^{\frac{1}{2}}). \tag{9}$$

5 Dataset

The dataset is created from the images of the patients acquired in the Department of Dermatology and Sexually Transmitted Disease at Katihar Medical College and Hospital, Bihar India[1]. The approval for data collection is given by Katihar Medical College, Institute ethics committee, IEC No. KMC/IEC/2013-2016/008/MD (Derma). The patients were treated in the outpatient department or admitted to the ward. These patients' images were categorized into the classes as per the diagnosis. The sample images of the patients and the extracted images are shown in Table 1. We extracted 200 skin lesions from four different categories of the diagnosis. The categories are Herpes zoster, herpes simplex, molluscum contagiosum and non-viral disorders. We extracted 50 skin lesions from different patients in image format for each category and resized them into 512 $\times$ 512 pixels. The sample images of our dataset are shown in Fig. 1.

6 Methodology and Results

We performed multiple experiments using image augmentation techniques and deep models to create a deep classification model. For the experimentation, we split our dataset into train and test such that train images contain 35 images from each class while the remaining 15 images are assigned to the test set (three iterations). All the results are calculated on this test dataset of 15 images per class. Also, we perform primary geometric data augmentation techniques like flip and rotate. Furthermore, we explore wGAN [3] and CAE [32] to increase the training data size. After image augmentation, we fine-tuned existing deep models namely VGG16 (batch Size = 16, learning rate = 0.001) [27], ResNET-18 (batch Size = 16, learning rate = 0.0001), ResNET-50 (batch Size = 128, learning rate = 0.0001) [13], MobileNet-V2 (batch size = 32, learning rate = 0.001) [24] and EfficientNet-V2 (batch size = 32, learning rate = 0.01) [30] on the training dataset and evaluated them on the test set. Each deep learning model underwent training for 500 epochs, employing early stopping with patience of 20 epochs. The Adam optimizer and binary cross-entropy loss function were utilized during the training process. All the experiments are performed on NVIDIA RTX A4000 16GB GPU. All the deep models are pre-trained on the ImageNet-1k dataset and taken from the PyTorch library [22]. The outline of our methodology is given in Fig. 2.

We also calculate the FID distance between the original images and generated/transformed images (Table 2). The images created using geometrical transformations achieved the lowest FID score of 0.59. Notably, the FID between original images and wGAN-generated images is higher than that between original images and CAE-generated images. This suggests that CAE-generated images in our dataset bear a closer resemblance to the original images compared to wGAN-generated images which is subsequently reflected in the results.

[1] The dataset is available on request at niti.chikhale19@gmail.com or sushilsavant786@gmail.com.

Table 2. Fréchet Inception Distance.

Models	100	200	300	400	500
wGAN	9.91	9.75	9.81	9.83	9.73
CAE	2.50	2.66	2.71	2.85	2.30
Geometric Augmentation	0.59				

Table 3. Results of classification of deep models on original dataset.

Models	Accuracy	Precision	Recall	F1 score
VGG16	0.88 ± 0.02	0.89 ± 0.02	0.88 ± 0.02	0.88 ± 0.02
ResNET-18	0.89 ± 0.01	0.89 ± 0.01	0.89 ± 0.01	0.89 ± 0.01
ResNET-50	0.95 ± 0.03	0.95 ± 0.03	0.95 ± 0.03	0.95 ± 0.04
MobileNet-V2	0.89 ± 0.02	0.90 ± 0.02	0.89 ± 0.02	0.89 ± 0.02
EfficientNet-V2	0.74 ± 0.01	0.80 ± 0.03	0.74 ± 0.01	0.73 ± 0.02

Table 4. Classwise accuracy of the best model (ResNET-50) from Table 3.

Class	Herpes simplex	Herpes zoster	Molluscum contagiosum	Non-viral
Mean accuracy	0.900	0.960	0.920	0.850

6.1 Classification and Evaluation of Deep Models on Original Dataset

Experimental Setup: We train deep models on the training dataset (35 images per class) and evaluate on the test dataset.

Results: The results of this experiment are shown in the Table 3 demonstrates the highest performance by ResNET-50 model with an accuracy of 0.95 ± 0.03, precision of 0.95 ± 0.03, recall of 0.95 ± 0.03 and F1 score of 0.95 ± 0.04 with the classwise accuracy of this model shown in Table 4. MobileNet-V2 also performs well, particularly in terms of precision, while EfficientNet-V2 exhibits comparatively lower performance across all metrics.

6.2 Classification and Evaluation of Deep Models on the Geometric Augmented Dataset.

Experimental Setup: We expand the training dataset by employing rotation and flipping techniques, including horizontal and vertical flips, and rotations at 0, 90, 180, and 270° to keep minimum non-natural distortions to spatial relationships within the images. Each original image yields six geometrically augmented counterparts, resulting in 210 images per class. Additionally, we merge the 35 original training images with the augmented set, totaling 245 images per class. The fine-tuned models are subsequently evaluated using the test dataset.

Table 5. Results of classification of deep models on the geometric augmented dataset.

Models	Accuracy	Precision	Recall	F1 score
VGG16	0.88 ± 0.00	0.89 ± 0.00	0.88 ± 0.00	0.88 ± 0.00
ResNET-18	0.93 ± 0.02	0.93 ± 0.02	0.93 ± 0.02	0.93 ± 0.02
ResNET-50	0.97 ± 0.01	0.97 ± 0.01	0.97 ± 0.01	0.97 ± 0.01
MobileNet-V2	0.92 ± 0.02	0.93 ± 0.02	0.92 ± 0.02	0.92 ± 0.02
EfficientNet-V2	0.83 ± 0.03	0.86 ±0.02	0.83 ± 0.03	0.83 ± 0.03

Table 6. Classwise accuracy of the best model (ResNET-50) from Table 5.

Class	Herpes simplex	Herpes zoster	Molluscum contagiosum	Non-viral
Mean accuracy	0.920	0.920	0.980	0.930

Results: Table 5 shows that ResNet-50 demonstrates the highest performance across all metrics, with an accuracy of 0.97 ± 0.01, precision of 0.97± 0.01, recall of 0.97 ± 0.01, and F1 score of 0.95 ± 0.01. The class-wise accuracy of this model is shown in Table 6. ResNET-50 stands out as the top-performing model followed by ResNET-18 and MobileNet-V2. VGG16 also demonstrates strong performance, albeit slightly lower than the aforementioned models. Conversely, EfficientNet-V2 shows slightly lower performance across all metrics compared to the other models.

6.3 Classification and Evaluation of Deep Models on wGAN Generated Images

Experimental Setup: We utilized wGAN to generate images from our initial training dataset, systematically escalating the quantity of generated images (Fig. 3). We established five distinct sets, each generating 100, 200, 300, 400, and 500 images per class, leveraging the initial training set containing 35 images per class. Moreover, we augmented the generated images by incorporating the original training dataset.

Results: Table 7 shows the results of models trained on wGAN-generated images. Here we can see that as we increase the number of images the performance of VGG16, ResNET-18 and EfficientNetV2 decreases except for the MobileNetV2. Table 8 shows the class-wise accuracy for the overall best-performing model i.e. ResNET50 for 300 images per class.

6.4 Classification and Evaluation of Deep Models on CAE Generated Images

Experimental Setup: We employed CAE to generate images from our original training dataset, systematically increasing the generated images (Fig. 4). We created five separate sets, each generating 100, 200, 300, 400, and 500 images per class, using the initial training set, which originally contained 35 images per class. Additionally, we augmented the generated images by integrating them with the original training dataset.

Table 7. Results of classification of deep models on the wGAN generated images.

Metrics	A	P	R	F1	A	P	R	F1	A	P	R	F1	A	P	R	F1	A	P	R	F1
Models	100				200				300				400				500			
V16	0.78 ± 0.01	0.78 ± 0.01	0.78 ± 0.01	0.78 ± 0.01	0.78 ± 0.01	0.78 ± 0.01	0.78 ± 0.01	0.78 ± 0.01	0.77 ± 0.01	0.78 ± 0.01	0.77 ± 0.01	0.77 ± 0.01	0.74 ± 0.02	0.74 ± 0.01	0.74 ± 0.02	0.74 ± 0.02	0.74 ± 0.01	0.74 ± 0.01	0.74 ± 0.01	0.74 ± 0.01
R18	0.93 ± 0.01	0.93 ± 0.01	0.93 ± 0.01	0.93 ± 0.01	0.90 ± 0.03	0.90 ± 0.02	0.90 ± 0.03	0.89 ± 0.03	0.91 ± 0.03	0.92 ± 0.03	0.91 ± 0.03	0.91 ± 0.03	0.90 ± 0.02	0.91 ± 0.02	0.90 ± 0.02	0.90 ± 0.02	0.92 ± 0.02	0.92 ± 0.02	0.92 ± 0.02	0.92 ± 0.02
R50	0.93 ± 0.01	0.93 ± 0.01	0.93 ± 0.01	0.93 ± 0.01	0.89 ± 0.01	0.89 ± 0.01	0.89 ± 0.01	0.89 ± 0.01	0.94 ± 0.02	0.94 ± 0.02	0.94 ± 0.02	0.94 ± 0.02	0.92 ± 0.02	0.93 ± 0.02	0.92 ± 0.02	0.92 ± 0.02	0.94 ± 0.02	0.94 ± 0.02	0.94 ± 0.02	0.94 ± 0.02
MV2	0.87 ± 0.01	0.88 ± 0.01	0.87 ± 0.01	0.87 ± 0.01	0.89 ± 0.04	0.89 ± 0.04	0.89 ± 0.04	0.89 ± 0.04	0.91 ± 0.04	0.91 ± 0.04	0.91 ± 0.04	0.91 ± 0.04	0.92 ± 0.05	0.92 ± 0.04	0.92 ± 0.05	0.92 ± 0.04	0.92 ± 0.03	0.93 ± 0.03	0.93 ± 0.03	0.93 ± 0.03
EV2	0.91 ± 0.02	0.91 ± 0.02	0.91 ± 0.02	0.90 ± 0.02	0.84 ± 0.01	0.87 ± 0.02	0.84 ± 0.01	0.84 ± 0.01	0.89 ± 0.03	0.90 ± 0.04	0.90 ± 0.03	0.90 ± 0.03	0.85 ± 0.04	0.86 ± 0.04	0.85 ± 0.04	0.84 ± 0.05	0.87 ± 0.04	0.87 ± 0.03	0.87 ± 0.04	0.87 ± 0.04

Here A - Accuracy, P - Precision, R - Recall, F1 - F1 score, V16-VGG16, R18 - ResNET-18, R50 - ResNET-50, MV2 - MobileNetV2 and EV2 - EfficientNetV2. 100, 200, 300, 400 and 500 are the number of images generated per class

Table 8. Classwise accuracy of the best model (ResNET-50) from Table 7.

Class	Herpes simplex	Herpes zoster	Molluscum contagiosum	Non-viral
Mean accuracy	0.880	0.940	0.900	0.780

Results: Table 9 shows the results of all models on CAE-generated images. We can see the increase in the performance of models as we increase the number of images. The ResNET-50 outperforms the rest of the models on the performance metrics. We also see an increase in the results of all models as we increase the number of images per class. Table 10 shows the class-wise accuracy of the ResNET50 model for 500 images.

Table 3 and 5 demonstrate an improvement in the results of the deep models after geometric augmentations. As we increase the training dataset, the results of deep models show improvement. These results show the utility of basic geometric augmentation techniques. We can also see the utility of the generative model namely wGAN and CAE. By comparing the results of Table 3, Table 7, and Table 9 we can see the improvement in the performance of deep models. Table 4, Table 6, Table 8 and Table 10 show the class-wise accuracies of outperforming models for respective experiments. We can see the increase in class-wise accuracy for normal dataset and geometric augmented dataset from Table 4 and Table 6. These results also show that the ResNET50 is able to classify all four classes without any class biases. After the wGAN image augmentation, the class-wise accuracy of the best model for the experiment is shown in Table 8. Here we can see the poor accuracy in Non-viral class and Herpes simplex suggesting poor image

Table 9. Results of classification of deep models on the CAE generated images.

Metrics	A	P	R	F1	A	P	R	F1	A	P	R	F1	A	P	R	F1	A	P	R	F1
Models	100				200				300				400				500			
V16	0.88 ± 0.01	0.88 ± 0.01	0.88 ± 0.01	0.88 ± 0.01	0.90 ± 0.00	0.90 ± 0.01	0.90 ± 0.00	0.90 ± 0.00	0.83 ± 0.02	0.85 ± 0.02	0.83 ± 0.02	0.84 ± 0.02	0.86 ± 0.02	0.86 ± 0.01	0.86 ± 0.02	0.86 ± 0.02	0.86 ± 0.01	0.87 ± 0.01	0.86 ± 0.01	0.86 ± 0.01
R18	0.90 ± 0.02	0.91 ± 0.01	0.90 ± 0.02	0.90 ± 0.02	0.92 ± 0.03	0.92 ± 0.03	0.92 ± 0.03	0.91 ± 0.03	0.92 ± 0.04	0.93 ± 0.04	0.93 ± 0.04	0.93 ± 0.04	0.93 ± 0.02	0.94 ± 0.02	0.93 ± 0.02	0.93 ± 0.02	0.92 ± 0.02	0.93 ± 0.02	0.92 ± 0.02	0.92 ± 0.02
R50	0.92 ± 0.03	0.93 ± 0.02	0.92 ± 0.03	0.92 ± 0.03	0.92 ± 0.02	0.93 ± 0.01	0.92 ± 0.02	0.91 ± 0.02	0.93 ± 0.02	0.94 ± 0.01	0.93 ± 0.02	0.93 ± 0.02	0.95 ± 0.02	0.95 ± 0.02	0.95 ± 0.02	0.95 ± 0.02	0.95 ± 0.02	0.95 ± 0.02	0.95 ± 0.02	0.95 ± 0.02
MV2	0.89 ± 0.02	0.89 ± 0.02	0.89 ± 0.02	0.89 ± 0.02	0.92 ± 0.04	0.93 ± 0.03	0.92 ± 0.04	0.92 ± 0.04	0.91 ± 0.03	0.92 ± 0.03	0.91 ± 0.03	0.91 ± 0.03	0.91 ± 0.02	0.92 ± 0.02	0.91 ± 0.02	0.91 ± 0.02	0.93 ± 0.03	0.93 ± 0.03	0.93 ± 0.03	0.93 ± 0.03
EV2	0.79 ± 0.08	0.84 ± 0.05	0.79 ± 0.08	0.78 ± 0.09	0.76 ± 0.06	0.80 ± 0.06	0.76 ± 0.06	0.75 ± 0.06	0.91 ± 0.03	0.91 ± 0.03	0.91 ± 0.03	0.91 ± 0.03	0.81 ± 0.08	0.83 ± 0.06	0.81 ± 0.08	0.81 ± 0.08	0.85 ± 0.04	0.88 ± 0.03	0.85 ± 0.04	0.85 ± 0.04

Here A - Accuracy, P - Precision, R - Recall, F1 - F1 score, V16- VGG16, R-18 - ResNET-18, R50 - ResNET-50, MV2 - MobileNetV2 and EV2 - EfficientNetV2. 100, 200, 300, 400 and 500 are the number of images generated per class

Table 10. Classwise accuracy of the best model (ResNET-50) from Table 9.

Class	Herpes simplex	Herpes zoster	Molluscum contagiosum	Non-viral
Mean accuracy	0.880	0.940	0.870	0.850

generation by wGAN. While in Table 10 shows the best models for CAE-generated images showing better class-wise accuracy than wGAN-generated images.

The VGG16 is a smaller 16-layer model while ResNET18 and ResNET50 are 18-layered and 50-layered deep models respectively [13,27]. ResNET18 and ResNET50 use residual blocks that allow gradients to flow through the network directly, addressing the vanishing gradient problem. The MobileNETv2 model is aimed at mobile and embedded vision applications, focusing on efficiency and low computational cost, while the EfficientNETv2 compound scaling method is used to balance network width, depth, and resolution [24,30]. Our results show that the performance of the VGG16 model is consistently low throughout. The geometric augmentation techniques are also not able to improve the performance of VGG16 do to its comparatively poor architecture, complex and small image dataset.

However, while comparing the performance of geometric-based augmentation to the generative model from Table 5, Table 7 and Table 9 our experiment shows the model trained on the geometric augmentations demonstrates better results. This in line with Baisi et al. [8] who demonstrate that different augmentation methods variably impact classification accuracy in skin lesions. This can be justified using the FID distance.

The wGAN and CAE models incorporate gaussian noise into the existing distribution to generate images. In contrast, transformation techniques typically do not introduce noise as part of the image generation process, as seen in Table 2. Also, the training data required for wGAN and CAE is more substantial than the data we used. Due to this, the noise introduced is higher, showing high FID in Table 2 suggesting a noticeable gap between the generated images and the real images in the feature space especially with very little training data for wGANs and CAE-driven augmentation.

7 Conclusion

Dermatological diseases are very common in the world and cause mental stigma due to disfigurement. The use of deep learning models can be helpful to medical experts for diagnosis/ detection purposes. In this study, we present a skin lesion dataset for the Herpes simplex, Herpes zoster, Molluscum contagiosum and Non-viral disorders. We train multiple deep-learning models on our dataset with different augmentation techniques. Also, we compare the performance of several augmentation techniques. Our study shows that basic augmentation techniques such as rotate, flip and mirror can be successfully used to increase the performance of deep models trained on smaller dataset. In Table 5 we can see a gradual increase in the results metrics than the result metrics of Table 3. Also, images generated from smaller datasets using wGAN and CAE may degrade the performance of the deep models. One reason behind this is that the FID distance between the generated and original images is on the higher side. Nevertheless, our study represents a step towards leveraging deep learning technologies for diagnosis and management of cutaneous viral diseases, with the potential to optimistically impact patient outcomes and public health initiatives.

References

1. Al-Rasheed, A., Ksibi, A., Ayadi, M., Alzahrani, A.I., Mamun Elahi, M.: An ensemble of transfer learning models for the prediction of skin lesions with conditional generative adversarial networks. Contrast Media Mol. Imaging **2023**, 1–15 (2023)
2. Arjovsky, M., Bottou, L.: Towards principled methods for training generative adversarial networks. arXiv preprint arXiv:1701.04862 (2017)
3. Arjovsky, M., Chintala, S., Bottou, L.: Wasserstein generative adversarial networks. In: Precup, D., Teh, Y.W. (eds.) Proceedings of the 34th International Conference on Machine Learning, vol. 70, pp. 214–223 (2017). https://proceedings.mlr.press/v70/arjovsky17a.html
4. Back, S., et al.: Robust skin disease classification by distilling deep neural network ensemble for the mobile diagnosis of herpes zoster. IEEE Access 20156–20169 (2021). https://doi.org/10.1109/ACCESS.2021.3054403
5. Bengio, Y., Yao, L., Alain, G., Vincent, P.: Generalized denoising auto-encoders as generative models. In: Advances in Neural Information Processing Systems, vol. 26 (2013)
6. Biau, G., Sangnier, M., Tanielian, U.: Some theoretical insights into Wasserstein GANs. J. Mach. Learn. Res. **22**(1), 5287–5331 (2021)
7. Codella, N.C.F., 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). IEEE (2018). https://doi.org/10.1109/ISBI.2018.8363547

8. Di Biasi, L., De Marco, F., Auriemma Citarella, A., Castrillón-Santana, M., Barra, P., Tortora, G.: Refactoring and performance analysis of the main CNN architectures: using false negative rate minimization to solve the clinical images melanoma detection problem. BMC Bioinform. **24**(1), 386 (2023)
9. Germain, N., et al.: Stigma in visible skin diseases - a literature review and development of a conceptual model. J. Eur. Acad. Dermatology Venereol. **35**(7) (2021). https://doi.org/10.1111/jdv.17110
10. Giotis, I., Molders, N., Land, S., Biehl, M., Jonkman, M.F., Petkov, N.: MED-NODE: a computer-assisted melanoma diagnosis system using non-dermoscopic images. Expert Syst. Appl. **42**(19), 6578–6585 (2015). https://doi.org/10.1016/j.eswa.2015.04.034
11. Goodfellow, I., et al.: Generative adversarial nets. In: Advances in Neural Information Processing Systems, vol. 27 (2014)
12. Han, S.S., Kim, M.S., Lim, W., Park, G.H., Park, I., Chang, S.E.: Classification of the clinical images for benign and malignant cutaneous tumors using a deep learning algorithm. J. Investig. Dermatol. **138**(7), 1529–1538 (2018). https://doi.org/10.1016/j.jid.2018.01.028
13. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 770–778 (2016). https://doi.org/10.1109/CVPR.2016.90
14. Heenaye-Mamode Khan, M., et al.: Multi-class skin problem classification using deep generative adversarial network (DGAN). Computat. Intell. Neurosci. **2022**, 1–13 (2022). https://doi.org/10.1155/2022/1797471
15. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: GANs trained by a two time-scale update rule converge to a local Nash equilibrium. In: Advances in Neural Information Processing Systems, vol. 30 (2017)
16. Hou, B., Yan, R.: Convolutional auto-encoder based deep feature learning for finger-vein verification. In: 2018 IEEE International Symposium on Medical Measurements and Applications (MeMeA), pp. 1–5. IEEE Press (2018). https://doi.org/10.1109/MeMeA.2018.8438719
17. Kawahara, J., Daneshvar, S., Argenziano, G., Hamarneh, G.: Seven-point checklist and skin lesion classification using multitask multimodal neural nets. IEEE J. Biomed. Health Inform. **23**(2), 538–546 (2019). https://doi.org/10.1109/JBHI.2018.2824327. https://ieeexplore.ieee.org/document/8333693/
18. Li, H., Pan, Y., Zhao, J., Zhang, L.: Skin disease diagnosis with deep learning: a review. Neurocomputing **464**, 364–393 (2021). https://doi.org/10.1016/j.neucom.2021.08.096
19. Muhaba, K.A., Dese, K., Aga, T.M., Zewdu, F.T., Simegn, G.L.: Automatic skin disease diagnosis using deep learning from clinical image and patient information. Skin Health Dis. **2**(1), e81 (2022). https://doi.org/10.1002/ski2.81
20. Mukherkjee, D., Saha, P., Kaplun, D., Sinitca, A., Sarkar, R.: Brain tumor image generation using an aggregation of GAN models with style transfer. Sci. Rep. **12**(1), 1–16 (2022)
21. Nigat, T.D., Sitote, T.M., Gedefaw, B.M.: Fungal skin disease classification using the convolutional neural network. J. Healthcare Eng. **2023**, 1–9 (2023). https://doi.org/10.1155/2023/6370416
22. Paszke, A., et al.: Contributors: Pytorch: an imperative style, high-performance deep learning library. In: Advances in Neural Information Processing Systems, vol. 32, pp. 8024–8035 (2019)
23. Pathak, A.K.: Study of drug utilization pattern for skin diseases in dermatology OPD of an Indian tertiary care hospital - a prescription survey. J. Clin. Diagn. Res. (2016). https://doi.org/10.7860/JCDR/2016/17209.7270
24. Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., Chen, L.C.: Mobilenetv2: inverted residuals and linear bottlenecks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4510–4520 (2018)

25. Seitzer, M.: PyTorch-fid: FID Score for PyTorch (2020). https://github.com/mseitzer/pytorch-fid, version 0.3.0
26. Shaik, R., Bodhapati, S.K., Uddandam, A., Krupal, L., Sengupta, J.: A deep learning model that diagnosis skin diseases and recommends medication. In: 2022 1st International Conference on the Paradigm Shifts in Communication, Embedded Systems, Machine Learning and Signal Processing (PCEMS), Nagpur, India, pp. 7–10. IEEE (2022). https://doi.org/10.1109/PCEMS55161.2022.9808065
27. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)
28. Sonawane, M.M., Gore, R.D., Gawali, B.W., Manza, R.R., Mendhekar, S.N.: Identification of skin disease using machine learning. In: Proceedings of the First International Conference on Advances in Computer Vision and Artificial Intelligence Technologies (ACVAIT 2022), pp. 99–113. Atlantis Press International BV, Dordrecht (2023). https://doi.org/10.2991/978-94-6463-196-8_9
29. Sreekala, K., et al.: Skin diseases classification using hybrid AI based localization approach. Comput. Intell. Neurosci. **2022**, 1–7 (2022). https://doi.org/10.1155/2022/6138490
30. Tan, M., Le, Q.: Efficientnetv2: smaller models and faster training. In: International Conference on Machine Learning, pp. 10096–10106. PMLR (2021)
31. Toğaçar, M., Cömert, Z., Ergen, B.: Intelligent skin cancer detection applying autoencoder, MobileNetV2 and spiking neural networks. Chaos Solitons Fractals **144**, 110714 (2021). https://doi.org/10.1016/j.chaos.2021.110714
32. Vincent, P., Larochelle, H., Bengio, Y., Manzagol, P.A.: Extracting and composing robust features with denoising autoencoders. In: Proceedings of the 25th International Conference on Machine Learning, ICML 2008, Helsinki, Finland, pp. 1096–1103. ACM Press (2008). https://doi.org/10.1145/1390156.1390294. http://portal.acm.org/citation.cfm?doid=1390156.1390294
33. Viola, P., Jones, M.: Rapid object detection using a boosted cascade of simple features. In: Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR 2001, vol. 1, p. I (2001). https://doi.org/10.1109/CVPR.2001.990517
34. Yang, S., Xiao, W., Zhang, M., Guo, S., Zhao, J., Shen, F.: Image data augmentation for deep learning: a survey. arXiv preprint arXiv:2204.08610 (2022)
35. Yi, X., Walia, E., Babyn, P.: Unsupervised and semi-supervised learning with categorical generative adversarial networks assisted by Wasserstein distance for dermoscopy image classification. arXiv preprint arXiv:1804.03700 (2018)
36. Zhang, Y.: A better autoencoder for image: convolutional autoencoder. In: ICONIP17-DCEC (2018). http://users.cecs.anu.edu.au/Tom.Gedeon/conf/ABCs2018/paper/ABCs2018_paper_58.pdf. Accessed 23 Mar 2017

FNOReg: Resolution-Robust Medical Image Registration Method Based on Fourier Neural Operator

Nikita A. Drozdov and Dmitry V. Sorokin(✉)

Laboratory of Mathematical Methods of Image Processing, Faculty of Computational Mathematics and Cybernetics, Lomonosov Moscow State University, Moscow, Russia
drozdovna@my.msu.ru, dsorokin@cs.msu.ru

Abstract. Medical image registration plays a crucial role in diagnosis, treatment planning, and anatomical studies. Classical methods, relying on iterative optimization algorithms, are complex and computationally intensive. Recent advances in deep learning, particularly with convolutional neural networks (CNNs) like VoxelMorph, have shown promise. However, they often yield non-smooth deformation fields and require inference at the same image resolution as the training data. To overcome these challenges, we introduce FNOReg, a novel model based on Fourier Neural Operators (FNOs), which can be trained on reduced-resolution images without quality loss and produces smoother deformation fields. We evaluated FNOReg on 2D and 3D datasets, demonstrating comparable quality to popular models like VoxelMorph, Fourier-Net, and TransMorph when trained at the same resolution. However, these models exhibit significant quality decreases of up to 24.9% for 2D and 24.6% for 3D data, when trained at halved resolutions. In contrast, FNOReg demonstrates only marginal quality decreases of up to 0.8% for 2D and 2.7% for 3D data. This flexibility is essential for efficiently handling large image resolutions, particularly in 3D imaging. Moreover, FNOReg produces smoother deformation fields. The code is available at https://github.com/anac0der/fnoreg.

Keywords: Biomedical image registration · Fourier Neural Operator · unsupervised learning

1 Introduction

Image registration is one of the key tasks in the field of image processing and is widely used in the analysis of medical images. Among the many applications of image registration, one can distinguish: (i) Combining information obtained using various imaging devices or protocols to facilitate diagnosis and treatment

The work was financially supported by the Russian Science Foundation under the research project No. 22-41-02002.

A. Antonacopoulos et al. (Eds.): ICPR 2024, LNCS 15313, pp. 163–177, 2025.
https://doi.org/10.1007/978-3-031-78201-5_11

planning. (ii) Studies that examine structural or anatomical changes in the same areas of the body. Registration can be classified based on various factors, including the dimensionality of the data (for example, 2D, 3D), the modality of the image (for example, computer tomography or MRI), objects in the image (for example, brain, lungs, and heart), and the form of transformation (rigid or non-rigid registration). The goal of algorithms for non-rigid registration is to predict a deformation field for the registered (moved) image so that the difference between the fixed and moved images is minimal. Despite the simple idea, the task turns out to be ill-posed and, therefore, difficult in practice. The solution to such a problem may not be unique [19]. In practice, some constraints are imposed on the deformation field, such as smoothness, symmetry, or diffeomorphism. These constraints serve as regularization for an iterative algorithm that solves the problem of minimizing the difference between images.

Before the deep learning era, non-rigid registration was performed using iterative optimization algorithms. Methods based on this approach include Free-Form Deformation [22], LDDMM [5], Demons [23], Elastix [13], ANT [3], NiftyReg [18], and Flash [25]. These methods are widely used and have a mathematical foundation but also require the selection of parameters for each pair of images and significant computational resources, which limits their use for the registration of large images in real-time. Currently, there is an increase in the number of approaches to the registration of biomedical images based on deep learning. Their advantage is the fast model inference without the need to select parameters for each pair of images. The most effective and popular methods, such as VoxelMorph [4], predict the deformation field using a convolutional auto-encoder (for example, U-Net [21]). Such methods are faster than classical ones since they only require forward propagation of a pair of images through the neural network and can be run on GPU out of the box. After the success of VoxelMorph, other neural network-based registration methods have appeared, in one way or another, improving the model which predicts the deformation field. For example, in [6], the authors replaced the convolutional blocks of the encoder with more complex transformer blocks with an attention mechanism, thus increasing model performance on medical datasets. However, neural network-based models often encounter issues with non-smooth deformation fields and require additional regularization in the loss function. A recent model called Fourier-Net [11] aims to address this problem by predicting a band-limited deformation field in the Fourier domain, followed by zero-padding and inverse Fourier transform. However, this approach involves low-pass filtering of the deformation field spectra with a rectangular step function, which can introduce various artifacts such as ringing and aliasing. Additionally, all current neural network-based approaches to image registration have a major disadvantage: they rely on classical convolutional layers with small kernels for deep learning, resulting in a degradation of registration quality when input data resolution changes.

Some classical approaches to registration are based on solving the Euler-Lagrange equation for the corresponding loss functional [5]. Typically, partial differential equations (PDEs) are solved using numerical techniques, but in recent

years, physics-informed neural operators (PINOs) [15] are widely used for such tasks. One of the PINO architectures is called the Fourier Neural Operator (FNO) [14] and is designed to solve the parametric PDE. FNO has one important property – since all operations are performed on a fixed set of frequencies in the Fourier space, this architecture is capable of modeling non-local convolution kernels. Consequently, FNO is weakly sensitive to the shape of input data and can be used for image registration at different resolutions. Recently, FNO was used for other image analysis problems such as image segmentation [24] and classification [12].

In this paper, we propose FNOReg – a model for image registration based on the original FNO architecture. The main advantage of this approach is its robustness to input data resolution. Our model has two significant improvements over the original FNO: we incorporated feature extractors based on the Spectral Transform layer [7] and used additional residual connections in the Fourier layers. Additionally, we tailored the training pipeline for the image registration problem. We evaluated FNO and FNOReg on the OASIS-1 dataset [16] for 2D and 3D data, and our model achieved similar quality compared to popular models widely used for image registration (VoxelMorph [4], TransMorph [6], FourierNet [11]). However, the performance of our FNOReg models does not significantly degrade when trained and evaluated on data with reduced resolutions, whereas other models exhibit a significant decrease in registration quality under these conditions. This feature of the method is important for reducing memory consumption when working with large data and can be especially useful for 3D data. Moreover, we compared the spectra of deformations obtained from different models and demonstrated that FNO and FNOReg provide smoother deformation fields compared to other models.

2 Methodology

2.1 From Image Registration to a System of PDEs

Let us define a spatial domain $\Omega \subset \mathbb{R}^2$ (we consider the two-dimensional case for simplicity, but all calculations can be extended to the case of an arbitrary dimension). We define $\boldsymbol{x} = (x_1, x_2) \in \Omega$ to be an arbitrary point in domain Ω and $\boldsymbol{\phi}(\boldsymbol{x}) = (\phi_1(\boldsymbol{x}),\ \phi_2(\boldsymbol{x}))$ to be a displacement field representing the image deformation. Let $f(\boldsymbol{x})$ be a continuous function in Ω with continuous first order partial derivatives on Ω. This function corresponds to the fixed image while similarly introduced functions $m(\boldsymbol{x})$ and $m_{\boldsymbol{\phi}}(\boldsymbol{x})$ corresponds to the moving and moved images accordingly. Next, we demonstrate that the image registration problem can be reduced to a parametric system of partial differential equations that can be subsequently solved using neural operators.

Let $\Omega_p \subset \Omega$ be a finite set of points from Ω which is a pixel grid for discrete versions of fixed, moving and moved images. Then, the common loss functional for image registration task in discrete case can be written as:

$$L_{f,m}(\boldsymbol{\phi}) = \frac{1}{|\Omega_p|} \sum_{\boldsymbol{p} \in \Omega_p} D_{f,m_\phi}(\boldsymbol{p}) + \frac{\lambda}{|\Omega_p|} \sum_{\boldsymbol{p} \in \Omega_p} ||\nabla \boldsymbol{\phi}(\boldsymbol{p})||^2,$$

where $D_{f,m_\phi}(\boldsymbol{p})$ is a pixel-wise difference between two images (for example, $D_{f,m}(\boldsymbol{p}) = (f(\boldsymbol{p}) - m(\boldsymbol{p}))^2$), the second term is the regularization term, and λ is the regularization parameter.

In continuous case, we can rewrite the loss functional in an integral form as

$$\hat{L}_{f,m}(\boldsymbol{\phi}) = \int_\Omega \left[D_{f,m_\phi} + \lambda \left(\left[\frac{\partial \phi_1}{\partial x_1}\right]^2 + \left[\frac{\partial \phi_1}{\partial x_2}\right]^2 + \left[\frac{\partial \phi_2}{\partial x_1}\right]^2 + \left[\frac{\partial \phi_2}{\partial x_2}\right]^2 \right) \right] dx_1 dx_2.$$

Now, if we denote the expression under integral as $F_{f,m}(\boldsymbol{\phi})$, and the partial derivatives $\partial\phi_i/\partial x_j$ as $\phi_{i,j}$, we can write the Euler-Lagrange system [8] for $\hat{L}_{f,m}$:

$$\begin{aligned} &\frac{\partial F_{f,m}}{\partial \phi_1} - \frac{\partial}{\partial x_1}\left(\frac{\partial F_{f,m}}{\partial \phi_{1,1}}\right) - \frac{\partial}{\partial x_2}\left(\frac{\partial F_{f,m}}{\partial \phi_{1,2}}\right) = 0; \\ &\frac{\partial F_{f,m}}{\partial \phi_2} - \frac{\partial}{\partial x_1}\left(\frac{\partial F_{f,m}}{\partial \phi_{2,1}}\right) - \frac{\partial}{\partial x_2}\left(\frac{\partial F_{f,m}}{\partial \phi_{2,2}}\right) = 0. \end{aligned} \tag{1}$$

It can be shown that after transformations the system (1) can be expressed as:

$$\frac{\partial D_{f,m_\phi}}{\partial \phi_1} - 2\lambda \cdot \Delta\phi_1 = 0; \qquad \frac{\partial D_{f,m_\phi}}{\partial \phi_2} - 2\lambda \cdot \Delta\phi_2 = 0. \tag{2}$$

where $\Delta = \frac{\partial^2}{\partial x_1^2} + \frac{\partial^2}{\partial x_2^2}$ is the Laplace operator. Thus, if the displacement field $\boldsymbol{\phi}_{opt}$ is the point minimum for the $\hat{L}_{f,m}(\boldsymbol{\phi})$, then it is the solution of system (2). This system is parameterized by the fixed image f and moving image m and we can apply neural operators to approximate the solution operator of this equation.

2.2 Fourier Neural Operator

Fourier Neural Operator (FNO) is a deep learning model for solving parametric partial differential equations [14]. FNO does not require knowledge of the equation itself and based only on the provided data, which leads to the classical supervised or unsupervised learning task.

The architecture of FNO is inspired by the mathematical principles that lies behind the solution of some types of operator equations. More specifically, if we have an equation $Lu(\boldsymbol{x}) = f(\boldsymbol{x})$ where L is a linear differential operator and $\boldsymbol{x} \in \Omega \subset \mathbb{R}^n$, then the solution $u(\boldsymbol{x})$ of this equation is given by formula

$$u(\boldsymbol{x}) = \int_\Omega G(\boldsymbol{x}, \boldsymbol{s}) f(\boldsymbol{s})\, d\boldsymbol{s},$$

where $G(\boldsymbol{x}, \boldsymbol{s})$ is a Green function of operator L. As proposed in [1,14], $u(\boldsymbol{x})$ can be approximated using the following iterative process:

$$u_{t+1}(\boldsymbol{x}) := \sigma\left(W_t u_t(\boldsymbol{x}) + (K u_t)(\boldsymbol{x})\right),\ t = 0, \ldots, N-1, \tag{3}$$

where

$$(Ku_t)(\boldsymbol{x}) := \int_{\Omega} \kappa_t(\boldsymbol{x} - \boldsymbol{y}) u_t(\boldsymbol{y})\, d\boldsymbol{y}. \tag{4}$$

Here $u_t(\boldsymbol{x}) \in \mathbb{R}^{d_{u_t}}$ is the function of $\boldsymbol{x}$, $W_t \in \mathbb{R}^{d_{u_{t+1}} \times d_{u_t}}$ is a learnable linear transformation, $\sigma(\cdot)$ is a nonlinear activation function and $\kappa_t(\boldsymbol{x}-\boldsymbol{y})$ is a learnable kernel function of convolutional operator K. We can think about $u_t(\boldsymbol{x})$ as the output of network hidden layer with d_{u_t} data channels.

We can apply the convolution theorem to the operator K and parameterize it directly in Fourier space:

$$(Ku_t)(\boldsymbol{x}) = \mathcal{F}^{-1}\left(\mathcal{F}(\kappa) \cdot \mathcal{F}(u_t)\right) = \mathcal{F}^{-1}\left(R \cdot \mathcal{F}(u_t)\right), \tag{5}$$

where $R(\boldsymbol{k}) \in \mathbb{C}^{d_{u_{t+1}} \times d_{u_t}}$ is a learnable function in spectral domain, $\mathcal{F}$ is the Fourier transform operator, and $\boldsymbol{k} = (k_1, \ldots, k_n) \in \mathbb{N}^n$ corresponds to non-negative frequencies. To reduce the computational complexity of the model, $R(\boldsymbol{k})$ is nonzero only for $k_j \leq k_{max,j}$, $j = 1, \ldots, n$. This step also provides FNO model to be resolution-robust, because $k_{max,j}$ does not depend on input data shape and so the convolutional layer in FNO operates only on the fixed number of harmonics in Fourier domain. The model architecture is given in Fig. 1.

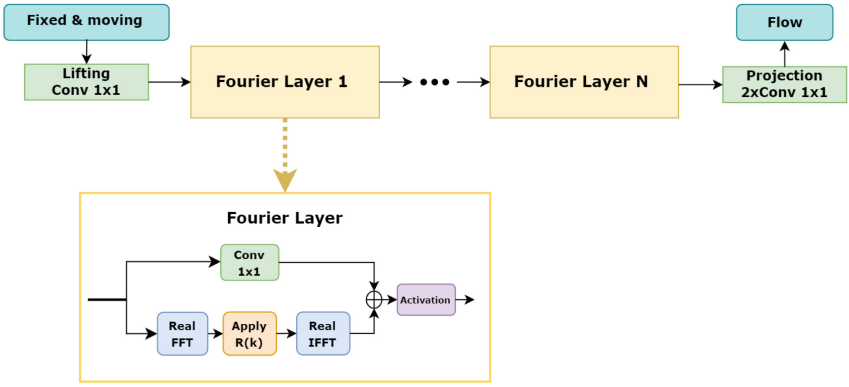


Fig. 1. FNO model architecture. Lifting layer maps the input data into a hidden space and increases the number of channels. Projection layer reduces the number of channel to get the output of the network. Fourier Layer consists the logic from Eq. 3.

2.3 FNOReg Model Architecture

The FNOReg architecture represents an enhancement of the classic FNO architecture designed for image registration. The significance of this improvement will be demonstrated later in Sect. 3.3. The model architecture is depicted in Fig. 2. In comparison to classic FNO [14], our model incorporates several enhancements

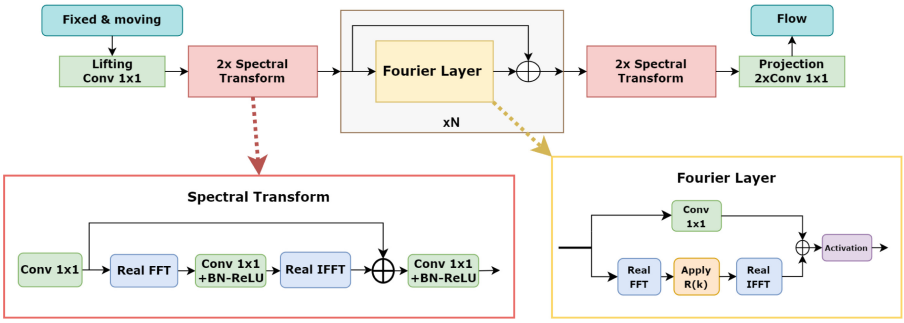


Fig. 2. FNOReg model architecture.

and additions that enable FNOReg to exhibit better quality and increased stability during training.

First, we utilize feature extractors (depicted as red rectangles in Fig. 2) designed with the Spectral Transform layer [7] instead of the standard convolutional layer. The main idea of the Spectral Transform is that convolution in the spectral domain possesses a global receptive field in the spatial domain, allowing us to extract non-local features from our images. This capability enables us to achieve higher registration quality without compromising the model's robustness to input resolution. Our research has shown that incorporating two Spectral Transform layers before and after the sequence of Fourier layers is enough for performance improvement without significantly complicating the model.

Our second improvement involves additional residual connections in the Fourier layers after the activation function. Residual connections tend to stabilize the learning process, and in our experiments FNOReg proved to be more robust when testing different model configurations than standard FNO.

The proposed improvements led to the better learning curve as can be seen in Fig. 3.

2.4 Training Pipeline

The process of training the model for image registration at each iteration includes a forward pass of data through the model, deforming the moving image according to the deformation field, calculating the loss function, and updating the model parameters. First, the fixed image f and moving image m are concatenated in the channel dimension, so the input data consist of an image with 2 channels. After propagating this image through the model, we obtain a deformation field ϕ. Then, we apply this deformation field to the moving image using a spatial transform layer [10], which is based on linear interpolation and implemented as described in [4]. Finally, we update the parameters of the model according to the value of the loss function, which is described below.

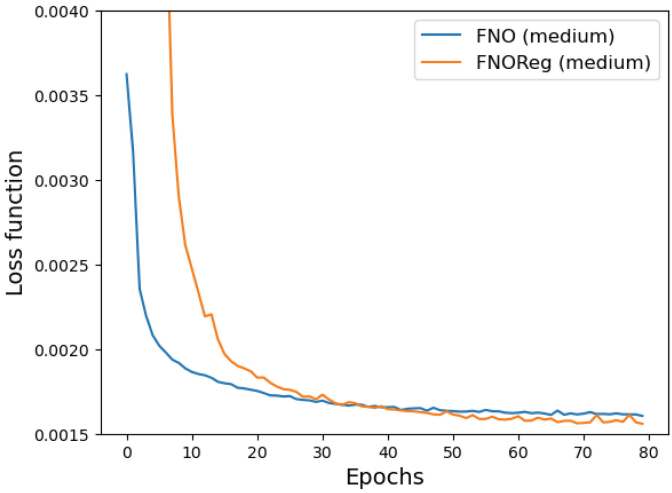


Fig. 3. The loss functions of the FNO and FNOReg models for the OASIS-1 2D dataset. The FNOReg loss starts at a value of 0.85, but its curve is much steeper than that of FNO, as evidenced by the curves intersecting.

Loss Function. We use a common loss function for image registration consisting of two components, $\mathcal{L}_{sim}$ and $\mathcal{L}_{smooth}$:

$$\mathcal{L}(f, m, \boldsymbol{\phi}) = \mathcal{L}_{sim}(f, m_{\boldsymbol{\phi}}) + \lambda \mathcal{L}_{smooth}(\boldsymbol{\phi}).$$

$\mathcal{L}_{sim}(f, m_{\boldsymbol{\phi}})$ penalizes the difference between the fixed and moved images. In our experiments, this component was mean squared error computed as:

$$MSE(f, m_{\phi}) = \frac{1}{|\Omega_p|} \sum_{\boldsymbol{p} \in \Omega_p} (f(\boldsymbol{p}) - m_{\phi}(\boldsymbol{p}))^2.$$

$\mathcal{L}_{smooth}$ is a regularization term that allows for more realistic and smoother deformation fields. We use a diffusion regularizer [2] implemented with finite difference approximation of gradient. Specifically, the following formula for $\mathcal{L}_{smooth}$ was used:

$$\mathcal{L}_{smooth}(\boldsymbol{\phi}) = \frac{1}{|\Omega_p|} \sum_{\boldsymbol{p} \in \Omega_p} ||\nabla \boldsymbol{\phi}(\boldsymbol{p})||^2.$$

For 3D data registration, we incorporate an additional term $\mathcal{L}_{seg}$ into the loss function to control segmentation overlap:

$$\mathcal{L}(f, m, s_f, s_{m_{\boldsymbol{\phi}}}, \boldsymbol{\phi}) = \mathcal{L}_{sim}(f, m_{\boldsymbol{\phi}}) + \lambda \mathcal{L}_{smooth}(\boldsymbol{\phi}) + \gamma \mathcal{L}_{seg}(s_f, s_{m_{\boldsymbol{\phi}}}),$$

where s_f and s_{m_ϕ} represent the anatomical segmentation of f and m_ϕ, respectively. We minimize the Dice loss [17] between s_f^k and $s_{m_\phi}^k$, where k denotes the k-th structure:

$$\mathcal{L}_{seg}(s_f, s_{m_\phi}) = 1 - \frac{2}{K} \sum_{k=1}^{K} \frac{|s_f^k \cap s_{m_\phi}^k|}{|s_f^k| + |s_{m_\phi}^k|}.$$

To enable automatic differentiation of the loss function, we follow a similar approach to [4], where s_f and s_{m_ϕ} are image volumes with K channels, each channel containing a binary mask defining the segmentation of a specific structure. The warped segmentation s_{m_ϕ} is computed using linear interpolation to ensure differentiability of the loss function.

3 Experiments

3.1 Datasets and Evaluation Metrics

OASIS-1 dataset [16] consists of brain MRI scans from 414 subjects. Each scan includes a segmentation of important anatomical areas. This dataset also provides 414 2D slices and their segmentation masks from corresponding 3D volumes. In our experiments, a preprocessed version of the 2D and 3D OASIS-1 dataset [9] was employed, where all 414 MRI scans were affinely aligned and cropped to the size of 160×192 and $160 \times 192 \times 224$, for 2D and 3D data respectively. For 2D data, we split all scans into 201 for training, 12 for validation, and 201 for testing. After pairing, we obtained 40200 pairs for training, 22 for validation, and 400 for testing. For 3D data, we used 384 scans for training, 10 for validation and 20 for testing; after pairing we ended up with 766 training pairs, 9 validation pairs and 19 test pairs. During the evaluation on the test data, we computed the Dice score for every anatomical area on the fixed and warped images. The evaluation metric for a single pair of images was the average Dice score across all anatomical areas, while the evaluation metric for the entire test dataset was the average across the evaluation metrics for each pair.

3.2 Implementation Details

We implemented our method using the PyTorch [20] framework and neuralop package [14] which is an original implementation of Fourier Neural Operator.

The models were trained as follows:

- For 2D data, we trained our models for 80 epochs using the Adam optimizer with a learning rate of 10^{-4} and a batch size of 8. The parameter λ was set to 0.01.
- For 3D data, we trained our models for 500 epochs using the Adam optimizer with a learning rate of 10^{-4} and a batch size of 1. The parameters λ and γ were set to 0.01.

3.3 Results

We compared FNO and FNOReg with several recent baseline methods based on convolutional neural networks and transformers, which are widely used for biomedical image registration:

1. **Fourier-Net** [11]: A CNN-based model that predicts a band-limited deformation field in the Fourier domain and then obtains the final deformation after zero-padding and inverse DFT.
2. **VoxelMorph** [4]: Model for image registration with a U-Net-based architecture.
3. **VoxelMorph-Large**: A deeper version of VoxelMorph (number of convolution kernels increases in 2 times on each layer, added some convolutional layers between encoder and decoder).
4. **VoxelMorph-Huge** [6]: A customized VoxelMorph model with a comparable parameter size to that of TransMorph that was used in [6] paper.
5. **TransMorph** [6]: A model with a transformer-based encoder and convolutional decoder.

To study the robustness to image resolution, we trained and evaluated all models in two scenarios. In the first scenario, the models were trained and evaluated at full resolution. In the second scenario, the models were trained on resolution reduced by 2 (80×96 pixels for 2D data and $80 \times 96 \times 112$ voxels for 3D data), but evaluated at full resolution, as in the first scenario. For TransMorph, we used the original authors' implementation in [6], and due to technical constraints of the implementation, it was not possible to train this model on halved resolution.

Results on 2D OASIS-1 Data. Table 1 presents the comparison of baseline methods with FNO and FNOReg on the 2D OASIS-1 dataset [16]. We considered three configurations for FNO-based models: small ($d_{u_t} = 16$, $N = 6$, $\boldsymbol{k}_{max} = (40, 48)$), medium ($d_{u_t} = 32$, $N = 12$, $\boldsymbol{k}_{max} = (40, 48)$), and large ($d_{u_t} = 32$, $N = 12$, $\boldsymbol{k}_{max} = (60, 72)$). The best Dice score when training at full resolution is achieved by VoxelMorph-Large, but its quality loss with training resolution reduction is 22.01% (0.777 Dice score versus 0.606). Fourier-Net has fewer parameters than other baseline models (except classical VoxelMorph, which achieved a comparable to Fourier-Net performance with only 91000 parameters) and achieves a 0.757 and 0.761 Dice score (for models with 16 and 32 channels, respectively) in the first training scenario, losing about 20–25% of quality in the second scenario. At the same time, the results on full resolution for FNO-based models are comparable to the results of TransMorph and VoxelMorph-Huge (0.775 DSC for large FNOReg and the same result for TransMorph), but the maximum loss of quality for training on downsampled images for FNO and FNOReg is 0.67% and 0.78%, respectively. For 2D data, the mean inference times were from 0.006 to 0.014 s for baseline models, and from 0.011 to 0.021 s for FNO-based models.

Table 1. Comparison of the models on the 2D OASIS-1 dataset; 1x and 0.5x indicate training on full and halved resolution, respectively. The Dice score is used for comparison (the bigger the better). The best values in each column of each group (baseline and FNO-based models) are highlighted in bold. The change in Dice score when training on halved resolution compared to full resolution is indicated in the last column.

Model name	#params	Dice on 1x	Dice on 0.5x	Metric change on 0.5x
Initial (only affine)	-	0.544	0.544	-
Fourier-Net (16 channels)	1.4M	0.757	0.611	−19.29%
Fourier-Net (32 channels)	6.1M	0.761	0.571	−24.97%
VoxelMorph	91K	0.760	**0.715**	−5.92%
VoxelMorph-Large	5.8M	**0.777**	0.606	−22.01%
VoxelMorph-Huge	21.1M	0.771	0.698	−9.47%
TransMorph	31.0M	0.775	-	-
FNO (small)	1.3M	0.751	0.746	−0.67%
FNOReg (small)	1.3M	0.753	0.751	−0.27%
FNO (medium)	10.0M	0.766	0.762	−0.52%
FNOReg (medium)	10.0M	0.769	0.763	−0.78%
FNO (large)	22.8M	0.769	0.766	−0.39%
FNOReg (large)	22.8M	**0.775**	**0.772**	−0.39%

In Fig. 4, the deformation fields obtained by the compared models and their spectra are depicted. From the visualization of deformation fields and their spectra, it can be observed that VoxelMorph and TransMorph deformation fields contain high-frequency modes, corresponding to the vertical cross on the spectra. The FNO-based models and Fourier-Net produce smoother deformation fields. However, in Fourier-Net, the spectrum is explicitly filtered with a box-shaped low-pass filter inside the model, which theoretically produces some aliasing and ringing artifacts that may affect the quality of the deformation fields.

Results on 3D OASIS-1 Data. Table 2 presents a comparison of baseline methods with FNOReg on the 3D OASIS-1 dataset [16]. Due to computational constraints, we considered only one configuration of the FNOReg model ($d_{u_t} = 12$, $N = 15$, $\boldsymbol{k}_{max} = (40, 48, 56)$). For 3D data, obtaining a reasonable hyperparameter configuration to train the standard FNO models was not feasible, as the training process did not converge. This highlights the significance of the improvements presented in the proposed FNOReg model. While TransMorph achieved the best Dice score when trained at full resolution, it was not possible to train it on half resolution due to limitations in the authors' implementation. VoxelMorph-Large attained the second-best result, but its quality degraded by 24.68% (0.863 Dice score versus 0.650) with training resolution reduction. Fourier-Net exhibited a similar quality loss as for 2D data when trained on half resolution. In contrast, the standard VoxelMorph model demonstrated a more significant quality decrease when trained on reduced resolution for 3D data com-

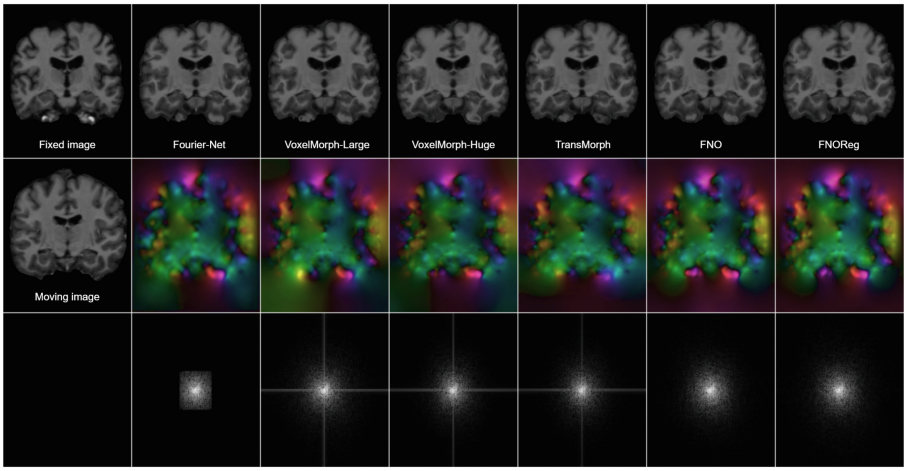


Fig. 4. Visual comparison of different methods on 2D OASIS-1 dataset. First column consists of fixed image and moving image. Other columns consist (from top to bottom) of moved image obtained from method, deformation field visualization, and deformation field DFT amplitude.

pared to 2D data. FNOReg also showed a slightly larger quality decrease when trained on downsampled images compared to 2D data (2.65%), but it remained considerably less than the other models. For 3D data, the mean inference times were from 0.087 to 1.048 s for baseline models, and 1.075 s for FNOReg.

Smoothness of Deformation Fields. To evaluate the smoothness of deformations obtained from different models, we calculate the percentage of voxels with a negative value of the Jacobian determinant of the deformation field (so-called "folded voxels" [6]) for each model. The results of this evaluation are presented in Table 3 for 2D data and in Table 4 for 3D data. As one can see, Fourier-Net achieves the best result at the original resolution due to the explicit low-pass filtering of the deformation field. However, in the case of Fourier-Net, the number of folded voxels strongly increases at halved resolution, leading to a degradation in performance. The smoothness of deformations from VoxelMorph and VoxelMorph-Huge slightly changes at halved resolution, but the absolute values of the metric at the original resolution for these models are worse than for other models (especially on 3D data). The FNOReg model at full resolution achieves the second-best result on 2D data and the forth-best result on 3D data. Moreover, the number of folded voxels in deformations obtained by FNOReg increases only slightly at halved resolution. To summarize, the FNOReg model

Table 2. Comparison of the models on the 3D OASIS-1 dataset; 1x and 0.5x indicate training on full and halved resolution, respectively. The Dice score is used for comparison (the bigger the better). The best values in each column of each group (baseline and FNO-based models) are highlighted in bold. The change in Dice score when training on halved resolution compared to full resolution is indicated in the last column.

Model name	#params	Dice on 1x	Dice on 0.5x	Metric change on 0.5x
Initial (only affine)	-	0.572	0.572	-
Fourier-Net (16 channels)	4.5M	0.814	0.617	−24.20%
Fourier-Net (32 channels)	17.8M	0.822	0.619	−24.70%
VoxelMorph	274K	0.834	0.737	−11.63%
VoxelMorph-Large	15.2M	0.863	0.650	−24.68%
VoxelMorph-Huge	63.3M	0.849	**0.722**	−14.96%
TransMorph	46.8M	**0.867**	-	-
FNOReg	98.1M	**0.830**	**0.808**	−2.65%

has the best absolute values for the folded voxels metric among the methods that do not show significant degradation in the deformation field smoothness at halved resolution. In other words, it achieves a tradeoff between smoothness of deformation at the original resolution and only its minor degradation at reduced resolution.

Table 3. Comparison of the smoothness of deformations from different models on the 2D OASIS dataset. A lower metric value indicates better performance. The best values in each column of each group (baseline and FNO-based models) are highlighted in bold.

Model name	% of $\lvert J_\phi\rvert < 0$ on 1x	% of $\lvert J_\phi\rvert < 0$ on 0.5x
Fourier-Net (16 channels)	**0.668** ± 0.353	6.802 ± 1.499
Fourier-Net (32 channels)	0.751 ± 0.374	7.477 ± 1.573
VoxelMorph	0.717 ± 0.368	0.743 ± 0.365
VoxelMorph-Large	0.744 ± 0.371	7.213 ± 1.632
VoxelMorph-Huge	0.726 ± 0.366	**0.512** ± 0.264
TransMorph	0.704 ± 0.365	-
FNO (large)	0.718 ± 0.356	0.802 ± 0.389
FNOReg (large)	**0.690** ± 0.355	**0.767** ± 0.370

Table 4. Comparison of the smoothness of deformations from different models on the 3D OASIS dataset. A lower metric value indicates better performance. The best values in each column of each group (baseline and FNO-based models) are highlighted in bold.

Model name	% of $\lvert J_\phi \rvert < 0$ on 1x	% of $\lvert J_\phi \rvert < 0$ on 0.5x
Fourier-Net (16 channels)	**0.167** ± 0.054	3.365 ± 0.357
Fourier-Net (32 channels)	0.213 ± 0.064	2.687 ± 0.326
VoxelMorph	0.772 ± 0.127	0.735 ± 0.131
VoxelMorph-Large	0.203 ± 0.047	3.197 ± 0.391
VoxelMorph-Huge	0.726 ± 0.123	**0.717** ± 0.100
TransMorph	0.680 ± 0.118	-
FNOReg	**0.314** ± 0.063	**0.572** ± 0.080

4 Conclusion

This paper introduces FNOReg, a novel model for image registration that builds upon the original FNO architecture. Using the properties of Fourier neural operators, FNOReg offers enhanced robustness to input data resolution. Our model incorporates two significant improvements over the original FNO: the integration of feature extractors based on the Spectral Transform layer [7] and the utilization of additional residual connections in the Fourier layers. Moreover, we tailored the training pipeline specifically for the image registration problem. In addition to improving registration quality, our enhancements to the FNO architecture stabilize the training process and enable the model to be trained in cases where the standard FNO fails.

Through comprehensive evaluations on the OASIS-1 2D and 3D datasets [16], we found that FNOReg achieves comparable quality to popular models widely used for image registration, such as VoxelMorph [4], TransMorph [6], and FourierNet [11]. Notably, our FNOReg models maintain consistent performance even when trained on data with reduced resolutions, unlike other models that exhibit a significant decrease in registration quality. This aspect of our method is particularly advantageous for reducing memory consumption when handling large datasets and can be especially beneficial for 3D data. Additionally, our comparison of deformation spectra obtained from different models reveals that FNO and FNOReg yield smoother deformation fields compared to their counterparts.

Acknowledgements. For computational experiments, neural network training, and fine-tuning, we utilized the MSU-270 supercomputer of Lomonosov Moscow State University with Nvidia Tesla A100 80GB GPUs.

References

1. Anandkumar, A., et al.: Neural operator: graph kernel network for partial differential equations. In: ICLR 2020 Workshop on Integration of Deep Neural Models and Differential Equations (2020)

2. Anoshina, N.A., Sorokin, D.V.: CNN-based unsupervised registration of time-lapse microscopy image sequences. Int. Arch. Photogrammetry Remote Sens. Spat. Inf. Sci. **XLVIII-2/W3-2023**, 9–14 (2023). https://doi.org/10.5194/isprs-archives-XLVIII-2-W3-2023-9-2023. https://isprs-archives.copernicus.org/articles/XLVIII-2-W3-2023/9/2023/
3. Avants, B.B., Tustison, N.J., Song, G., Cook, P.A., Klein, A., Gee, J.C.: A reproducible evaluation of ants similarity metric performance in brain image registration. Neuroimage **54**(3), 2033–2044 (2011)
4. Balakrishnan, G., Zhao, A., Sabuncu, M.R., Guttag, J., Dalca, A.V.: Voxelmorph: a learning framework for deformable medical image registration. IEEE Trans. Med. Imaging **38**(8), 1788–1800 (2019)
5. Beg, M.F., Miller, M.I., Trouvé, A., Younes, L.: Computing large deformation metric mappings via geodesic flows of diffeomorphisms. Int. J. Comput. Vis. **61**, 139–157 (2005)
6. Chen, J., Frey, E.C., He, Y., Segars, W.P., Li, Y., Du, Y.: Transmorph: transformer for unsupervised medical image registration. Med. Image Anal. **82**, 102615 (2022)
7. Chi, L., Jiang, B., Mu, Y.: Fast Fourier convolution. In: Advances in Neural Information Processing Systems, vol. 33, pp. 4479–4488 (2020)
8. Courant, R., Hilbert, D.: Methods of mathematical physics, vol. I. Phys. Today **7**(5), 17 (1954)
9. Hoopes, A., Hoffmann, M., Greve, D.N., Fischl, B., Guttag, J., Dalca, A.V.: Learning the effect of registration hyperparameters with hypermorph. J. Mach. Learn. Biomed. Imaging **1** (2022)
10. Jaderberg, M., Simonyan, K., Zisserman, A., et al.: Spatial transformer networks. In: Advances in Neural Information Processing Systems, vol. 28 (2015)
11. Jia, X., et al.: Fourier-net: fast image registration with band-limited deformation. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 37, pp. 1015–1023 (2023)
12. Johnny, W., Brigido, H., Ladeira, M., Souza, J.C.F.: Fourier neural operator for image classification. In: 2022 17th Iberian Conference on Information Systems and Technologies (CISTI), pp. 1–6. IEEE (2022)
13. Klein, S., Staring, M., Murphy, K., Viergever, M.A., Pluim, J.P.: Elastix: a toolbox for intensity-based medical image registration. IEEE Trans. Med. Imaging **29**(1), 196–205 (2009)
14. Li, Z., et al.: Fourier neural operator for parametric partial differential equations. arXiv preprint arXiv:2010.08895 (2020)
15. Li, Z., et al.: Physics-informed neural operator for learning partial differential equations. ACM/JMS J. Data Sci. (2021)
16. Marcus, D.S., Wang, T.H., Parker, J., Csernansky, J.G., Morris, J.C., Buckner, R.L.: Open access series of imaging studies (OASIS): cross-sectional MRI data in young, middle aged, nondemented, and demented older adults. J. Cogn. Neurosci. **19**(9), 1498–1507 (2007)
17. Milletari, F., Navab, N., Ahmadi, S.A.: V-net: fully convolutional neural networks for volumetric medical image segmentation. In: 2016 Fourth International Conference on 3D Vision (3DV), pp. 565–571. IEEE (2016)
18. Modat, M., et al.: Fast free-form deformation using graphics processing units. Comput. Methods Programs Biomed. **98**(3), 278–284 (2010)
19. Modersitzki, J.: Numerical Methods for Image Registration. OUP Oxford (2003)
20. Paszke, A., et al.: PyTorch: an imperative style, high-performance deep learning library. In: Advances in Neural Information Processing Systems, vol. 32 (2019)

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. Rueckert, D., Sonoda, L.I., Hayes, C., Hill, D.L., Leach, M.O., Hawkes, D.J.: Non-rigid registration using free-form deformations: application to breast MR images. IEEE Trans. Med. Imaging **18**(8), 712–721 (1999)
23. Vercauteren, T., Pennec, X., Perchant, A., Ayache, N.: Diffeomorphic demons: efficient non-parametric image registration. Neuroimage **45**(1), S61–S72 (2009)
24. Wong, K.C., Wang, H., Syeda-Mahmood, T.: Fnoseg3D: resolution-robust 3D image segmentation with Fourier neural operator. In: 2023 IEEE 20th International Symposium on Biomedical Imaging (ISBI), pp. 1–5. IEEE (2023)
25. Zhang, M., Fletcher, P.T.: Fast diffeomorphic image registration via Fourier-approximated lie algebras. Int. J. Comput. Vis. **127**, 61–73 (2019)

Harmonized Spatial and Spectral Learning for Generalized Medical Image Segmentation

Vandan Gorade[1(✉)], Sparsh Mittal[2(✉)], Debesh Jha[1], Rekha Singhal[3], and Ulas Bagci[1(✉)]

[1] Machine and Hybrid Intelligence Lab, Northwestern University, Chicago, IL, USA
{vandan.gorade,debesh.jha,ulas.bagci}@northwestern.edu
[2] ECE Department, Indian Institute of Technology, Roorkee, Roorkee, India
sparsh.mittal@ece.iitr.ac.in
[3] TCS Research, Mumbai, India
rekha.singhal@tcs.com

Abstract. Deep learning has demonstrated remarkable achievements in medical image segmentation. However, prevailing deep learning models struggle with poor generalization due to (i) intra-class variations, where the same class appears differently in different samples, and (ii) inter-class independence, resulting in difficulties capturing intricate relationships between distinct objects, leading to higher false negative cases. This paper presents a novel approach that synergies spatial and spectral representations to enhance domain-generalized medical image segmentation. We introduce the innovative Spectral Correlation Coefficient objective to improve the model's capacity to capture middle-order features and contextual long-range dependencies. This objective complements traditional spatial objectives by incorporating valuable spectral information. Extensive experiments reveal that optimizing this objective with existing architectures like UNet and TransUNet significantly enhances generalization, interpretability, and noise robustness, producing more confident predictions. For instance, in cardiac segmentation, we observe a 0.81 pp and 1.63 pp (pp = percentage point) improvement in DSC over UNet and TransUNet, respectively. Our interpretability study demonstrates that, in most tasks, objectives optimized with UNet outperform even TransUNet by introducing global contextual information alongside local details. These findings underscore the versatility and effectiveness of our proposed method across diverse imaging modalities and medical domains. Code is available at https://github.com/vangorade/HarmonizedSS_ICPR2024.

1 Introduction

Medical image segmentation (MIS) is crucial for supporting clinicians in identifying injuries, monitoring diseases, and planning treatments. Deep learning models have allowed automated delineation of critical structures and organs, enhancing

A. Antonacopoulos et al. (Eds.): ICPR 2024, LNCS 15313, pp. 178–193, 2025.
https://doi.org/10.1007/978-3-031-78201-5_12

the precision and efficiency of treatment. However, existing deep learning models for MIS [10,12,15,16] lack generalization [18,23], i.e., they fail to accurately segment new and unseen data. The challenge to generalization includes the diversity in medical imaging data stemming from variations in imaging devices, protocols, patient demographics, and even the inherent biases [13,22,24,26] present in deep learning models. The diversity manifests as intra-class variations or inter-class independence. (1) Figure 1(a) depicts *intra-class variations*. It refers to the differences in appearance (size, shape, location, and texture) within a single class, such as organs like the stomach or polyps, across diverse samples from multiple acquisition equipment.

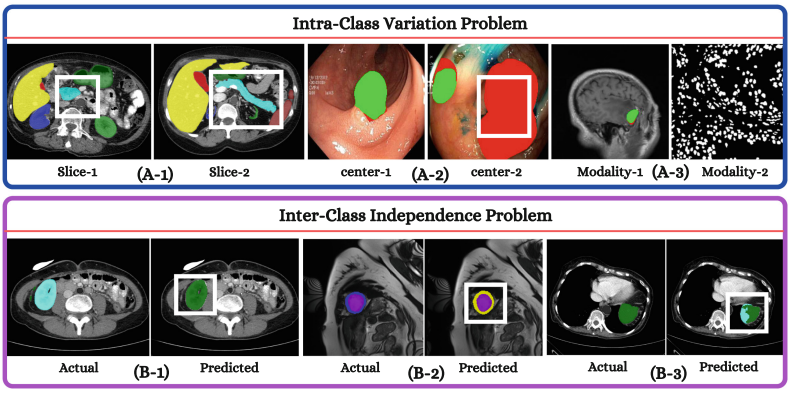


Fig. 1. (A-1) Appearance disparities within a single class of patient slices, highlighted by white bounding boxes indicating pancreas variation. (A-2) variation in ROI across data acquisition centers. (A-3) ROI variation between modalities. (B-1/2/3) Models face challenges in effectively capturing intricate inter-class relationships, as highlighted by the presence of white bounding boxes. These indicate instances of false negatives, a result of the model's struggle to learn relationships between classes effectively.

(2) Figure 1(b) shows the *inter-class independence*. It stems from the model's struggle to effectively model the intricate relationships between distinct objects or classes within the data. For instance, accurately segmenting multiple organs in a CT scan requires a deep understanding of their spatial interactions, influencing their appearances and boundaries. Disregarding such inter-class dependencies may lead to increased false negatives and poor generalization.

We introduce a novel approach that integrates prevalent spatial objectives, such as the Dice Similarity Coefficient, with an innovative objective termed the *Spectral Correlation Coefficient*. Unlike spatial objectives that concentrate on pixel-level comparisons, the Spectral Correlation Coefficient operates in the frequency domain. This integration is intended to augment segmentation models' effectiveness in apprehending middle-order features and contextual long-range dependencies. Both play a vital role in addressing variations within the

same class (intra-class variations) and establishing connections between different classes (inter-class dependencies). In contrast to previous methods [20,28,29], our approach is unique in that it avoids the prevailing practice of applying the Fast Fourier Transform (FFT) to input images. This novelty is important because applying FFT to input images can inadvertently restrict the model's ability to comprehend contextual relationships between objects due to the presence of ROI-irrelevant information in the images, as shown in Fig. 2.

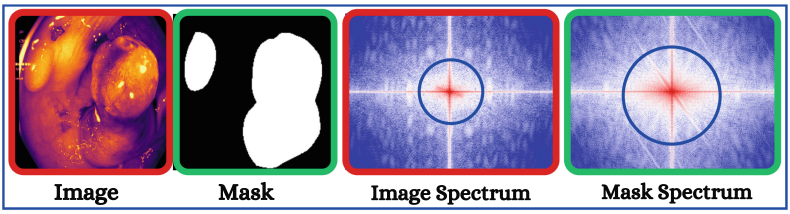


Fig. 2. A dense low-frequency spectrum (in the middle) indicates that the mask spectrum retains more object information than the image spectrum.

The Spectral Correlation Coefficient can reveal intricate patterns that remain hidden in the spatial domain. Its computation involves an $\mathcal{O}(N \log N)$ FFT operation, balancing performance with computational overhead. Our contributions are outlined as follows:

- We introduce a novel *Spectral Correlation Coefficient* objective, which integrates seamlessly with any architecture. It synergizes spatial and spectral representations, and enables effectively capturing middle-order features and long-range dependencies for domain-generalized MIS.
- We emphasize that addressing intra-class variations and establishing inter-class dependencies are crucial for achieving domain generalization in medical image segmentation.
- We conduct experiments on ***eight*** medical image datasets, comprising diverse imaging modalities and medical domains, e.g., including CT scans, MRIs, skin lesions, histopathology, and polyps. Our method demonstrates significant improvements in segmentation model out-of-distribution (OOD) robustness, enhancing generalization, interpretability, noise resilience, and calibration.

2 Proposed Method

2.1 Motivation

Middle-Order Features: Most current segmentation methods rely on spatial objectives to establish correspondence between predicted labels y and the ground truth $\hat{y}$. However, the raw pixels in the spatial domain exhibit significant noise and often encompass low-order statistics [5,27]. Transformers and

Convolutional Neural Networks (CNNs) possess distinct low-pass and high-pass filtering properties [14,24], respectively. However, both transformers and CNNs struggle to effectively model certain frequency bands, particularly those related to middle-order features.

Incorporating the medium frequency descriptor, such as the Histogram of Oriented Gradients (HOG), has proven beneficial in enhancing middle-order features [27]. This observation has prompted the hypothesis that gaining insights into medium frequencies could potentially aid the model in more effectively learning middle-order features. *Our proposition is that by comprehensively modeling these middle-order features, we can overcome the challenges posed by intra-class variations and inter-class independence.*

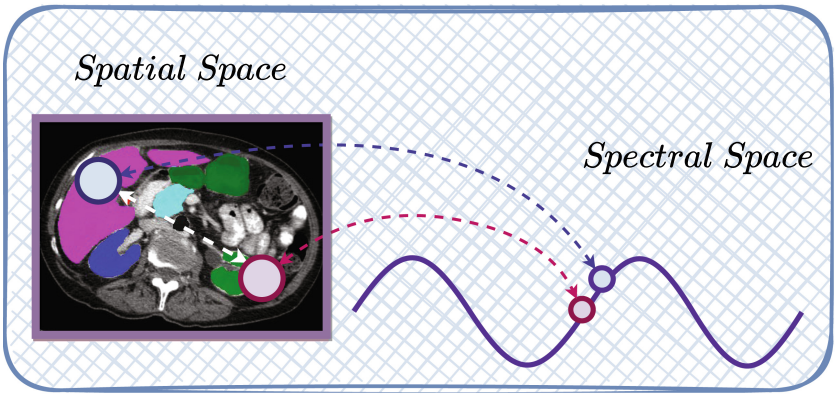


Fig. 3. Large variations in spatial space correspond to small variations in spectral space and vice versa.

Long-Range Dependencies: Existing CNN architectures face challenges in learning global features [25], which can lead to difficulties in capturing long-range dependencies. In contrast, transformers excel at modeling long-range dependencies [24]. Nevertheless, we have observed that solely learning long-range dependencies through random patch interactions does not suffice to grasp inter-class dependencies. *We propose that to effectively learn these inter-class dependencies, a model should focus on capturing long-range dependencies between pertinent regions rather than redundant ones.* The frequency space inherently facilitates the modeling of long-range dependencies because minor alterations in frequency space correspond to substantial spatial shifts, as demonstrated in Fig. 3. With the proposed spectral correlation coefficient, as a model learns correlations between the FFT mask and the predicted mask, it effectively learns the correlations among different frequency components. These components encapsulate only relevant class-related information, allowing us to capture and model inter-class dependencies effectively.

2.2 Problem Formulation

Medical image segmentation utilize a mapping function f, which assigns labels y to pixels x, where the inferred segmentation label is $\hat{y}$. The loss function typically combines Binary Cross Entropy (BCE) and the Dice similarity coefficient (Dice), which evaluate the correspondence between predicted labels y and ground truth segmentation $\hat{y}$:

$$\mathcal{L}_{\text{spatial}} = BCE(y, \hat{y}) + (1 - Dice(y, \hat{y})) \tag{1}$$

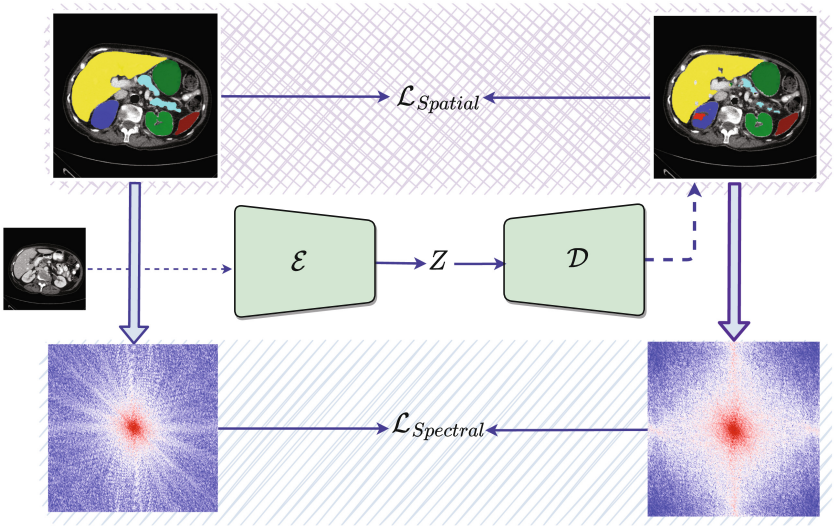


Fig. 4. Method Workflow: Starting with image x and mask y, an encoder-decoder network generates $\hat{y}$. Transforming to spectral space yields y_{freq} and $\hat{y}_{freq}$. Training involves spatial objective $\mathcal{L}_{spatial}$ between y and $\hat{y}$, alongside spectral objective $\mathcal{L}_{spectral}$ between y_{freq} and $\hat{y}_{freq}$.

Our goal is to augment f to transcend specific training domains and generalize effectively across diverse medical image datasets. This entails capturing common features and patterns across different domains. We introduce the Spectral Correlation Coefficient denoted as $\mathcal{L}_{\text{spectral}}$. This harmonizes the frequency components between predicted and ground-truth masks, effectively mitigating the limitations inherent in $\mathcal{L}_{\text{spatial}}$. Through the synergistic fusion of $\mathcal{L}_{\text{spatial}}$ and $\mathcal{L}_{\text{spectral}}$, the network can more effectively capture intricate inter-class relationships and intra-class variations. This collaborative approach bolsters the model's robustness and efficacy across diverse imaging scenarios. Figure 4 summarizes our approach.

2.3 Spectral Correlation Coefficient as Regularizer

Given two spatial binary masks, y and $\hat{y}$, we apply FFT to convert them to the frequency domain. This yields y_{freq} and $\hat{y}_{\text{freq}}$, which reveal the frequency components inherent to each signal. Then, we compute the complex inner product between y_{freq_i} and $\hat{y}_{\text{freq}_i}$ for each index i. This complex inner product encapsulates both amplitude and phase interactions in a singular value: $y_{\text{freq}_i} \cdot \overline{\hat{y}_{\text{freq}_i}}$. This helps elucidate the interplay among these frequency components.

By extracting the real component of this complex inner product, denoted as $\text{Re}(y_{\text{freq}_i} \cdot \overline{\hat{y}_{\text{freq}_i}})$, we can discern the interplay between the real and imaginary parts of these frequency components. This reveals the fundamental correlation between them. To measure the strength of these frequency components, we compute the squared magnitude (norm) of each frequency component, yielding $|y_{\text{freq}_i}|^2$ and $|\hat{y}_{\text{freq}_i}|^2$.

These insights culminate in $\mathcal{L}_{spectral}$, a quantitative metric for correlating y_{freq} and $\hat{y}_{\text{freq}}$:

$$\mathcal{L}_{\text{spectral}} = \frac{2\sum_{i=1}^{N}(\text{Re}(y_{\text{freq}_i})\text{Re}(\hat{y}_{\text{freq}_i}) + \text{Im}(y_{\text{freq}_i})\text{Im}(\hat{y}_{\text{freq}_i}))}{\sum_{i=1}^{N}(|y_{\text{freq}_i}|^2 + |\hat{y}_{\text{freq}_i}|^2)} \tag{2}$$

Here, N denotes the number of samples in the batch. This equation affords a comprehensive perspective on the similarity between y_{freq} and $\hat{y}_{\text{freq}}$, effectively encapsulating both their amplitude and phase characteristics. $\mathcal{L}_{spectral}$ stands as a vital metric for quantifying the correlation and shared attributes among frequency components across distinct signals. Our final loss function, $\mathcal{L}_{final} = \mathcal{L}_{spatial} + \lambda \times \mathcal{L}_{spectral}$. Synergizes the complementary representations of individual loss functions. Here, λ is a hyperparameter for smoothly interpolating between spatial and spectral representation. Please refer to supplementary material for sensitivity analysis of λ.

3 Experimental Platform

We conducted experiments on eight open-source MIS datasets to tackle diverse tasks spanning different anatomical structures. (1) The Synapse Multi-Organ Segmentation dataset [1] comprised 30 clinical CT cases, each equipped with annotated segmentation masks for eight distinct abdominal organs. We allocated 18 cases for training and 12 cases for testing [7]. (2) The ACDC dataset [2] has 100 cardiac MRI exams, with labels for the left ventricle (LV), right ventricle (RV), and myocardium (MYO). The train:validation:test split is 70:10:20 [7]. (3 & 4) For polyp segmentation, we used Kvasir-SEG [17] and PolypGen dataset [3]. Kvasir-SEG, containing 1000 images, was employed for training, with the official split of 880 training images and the remainder for testing. PolypGen has 1537 images. It assessed model performance under an out-of-distribution (OOD) setting. (5 & 6) For skin lesion segmentation, we use ISIC-18 [9] and ISIC-17 [9] datasets. We used the same split as the prior work [4,6]. The ISIC-17

test dataset [21], comprising 650 images, served for OOD testing. (7) For nuclei segmentation, we used the MoNuSeg dataset [19], which has 30 images for training and 14 for testing. (8) Brain Tumour Segmentation (BTSeg) dataset [8] has 3064 T1-weighted contrast-enhanced images, spanning three types of brain tumors with corresponding binary masks. The train: test split is roughly 80:20. For all datasets we set $\lambda = 0.2$. Please refer to supplementary material for sensitivity analysis of λ.

Metrics: We used the Dice Similarity Coefficient (DSC) and the 95th percentile Hausdorff Distance (HD) metrics on the Synapse and ACDC datasets. For the ISIC-18 and BTSeg datasets, we use Intersection over Union (IOU), DSC, Specificity (SP), Sensitivity (SE), and Accuracy (ACC). We also use the Expected/Mean Calibration Error (ECE/MCE) to assess the calibration. Lower HD, ECE, and MCE values are better, while higher values for other metrics are better.

Table 1. Results on the Kvasir-SEG, ISIC-18, MoNuSeg and BTSeg dataset.

Method	Kvasir-SEG				ISIC-18				MoNuSeg		BTSeg			
	DSC	IOU	SE	SP	DSC	IOU	SE	SP	DSC	IOU	DSC	IOU	SE	SP
UNet	88.77	76.97	81.55	**98.72**	91.53	80.34	**85.97**	96.65	72.85	58.80	84.40	68.94	74.58	**99.85**
TransUNet	87.68	75.56	81.94	98.17	90.92	79.42	86.30	95.44	76.92	63.01	83.15	67.69	73.63	99.84
UNet (Ours)	**89.40**	**77.03**	**83.19**	98.53	**91.74**	**80.51**	85.44	**96.93**	**73.38**	**58.97**	**85.48**	**69.91**	**75.33**	**99.85**
TransUNet(Ours)	**89.52**	**77.69**	**83.61**	**98.34**	**92.00**	**80.81**	**87.21**	**95.79**	**77.66**	**63.63**	**87.19**	**71.03**	**74.94**	**99.89**

Implementation Details: We used 224×224 images and train on RTX 2080 GPUs using Pytorch. During training, we used a batch size of 8 and a learning rate of 0.01. The encoder was initialized with weights pre-trained on ImageNet. We utilized the SGD optimizer with a momentum of 0.9 and weight decay of 0.0001. We employed data augmentations, such as flipping and rotating.

Techniques for Comparison: To ensure a comprehensive and fair evaluation, we chose (1) a CNN-based network, namely UNet with ResNet50 pretrained on ImageNet as the encoder. (2) a transformer-based network, namely TransUnet. It has a similar configuration as above, except that it has a transformer bottleneck with eight attention heads. We trained these models both with and without our proposed $\mathcal{L}_{spectral}$ regularization technique. We refer to UNet optimized using $\mathcal{L}_{spatial}$ as UNet, and the one optimized using $\mathcal{L}_{final}$ as UNet (ours); same for TransUNet and TransUNet (ours). We maintained uniformity in hyperparameters and architectural configurations across all the methods to isolate the effect of the proposed regularization technique.

4 Experimental Results

4.1 Robustness Against Intra-class Variations

We conducted an extensive analysis to assess the effectiveness of our proposed approach in addressing intra-class variation challenges. Table 1 presents a com-

prehensive comparison of methods, highlighting the outcomes of our study. Our proposed approach demonstrates clear advantages across diverse datasets that exhibit a wide range of anatomical variations. Notably, our method showcases the ability to accurately delineate both small and large anatomical structures while maintaining fine boundaries.

Results on Kvasir-SEG and ISIC-18: From a quantitative standpoint, on the Kvasir-SEG dataset, UNet(ours) performs comparably or better than the baselines. In fact, the improvements are even more pronounced with TransUNet, with increases of 1.84 pp, 2.13 pp, and 1.67 pp for DSC, IOU, and SE, respectively. The improvement in sensitivity indicates the model's ability to capture positive instances more effectively and reduce false negatives. The ISIC-18 dataset exhibits a similar trend, reaffirming the effectiveness of our approach. A qualitative analysis, as depicted in Fig. 5, supports our findings. Our proposed method can effectively capture intra-class variations. However, the performance of our method may depend on the nature of the dataset. For instance, datasets such as Kvasir-SEG and ISIC-18 predominantly include segmentation masks with single foreground objects. Such scenarios may limit the effectiveness of our method.

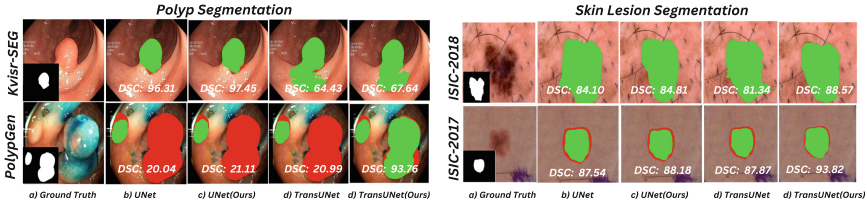


Fig. 5. Segmentation maps for polyp and skin lesion segmentation: Kvasir-SEG and ISIC-18 are trained under IID settings, while PolypGen and ISIC-17 are treated as OOD datasets. Actual and predicted pathological regions are shown in Red and Green, respectively. (Color figure online)

Results on MoNuSeg and BTSeg. Our approach improves the results on both these datasets (Table 1). Specifically, on the MoNuSeg dataset, we observe substantial advancements in performance for both UNet and TransUNet architectures, with increases of 0.53 pp and 0.58 pp in DSC, respectively. Similarly, on the BTSeg dataset, our method significantly benefits both UNet and TransUNet, showcasing DSC improvements of 1.08 pp and 4.04 pp, respectively. Figure 6 offers qualitative insights into our results. Particularly noteworthy is the considerable improvement TransUNet(ours) demonstrated over other baseline methods. This substantial improvement underscores the crucial role of capturing contextual long-range dependencies, achieved through our proposed $\mathcal{L}_{spectral}$ objective. This proves especially advantageous in scenarios like MoNuSeg and BTSeg, where segmentation tasks encompass a wide range of object variations in terms of size, shape, and spatial distribution.

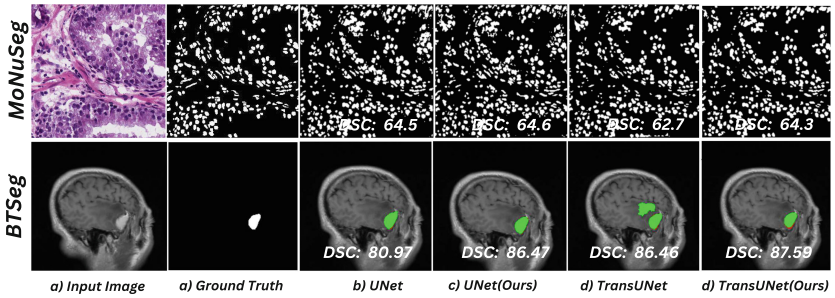


Fig. 6. Segmentation maps on MoNuSeg and BTSeg. Actual and predicted regions are shown in Red and Green, respectively. (Color figure online)

Table 2. Results on the Synapse and ACDC dataset. **Blue** indicates the best result.

	Synapse									ACDC			
Method	**Mean**	**Class-wise Dice Similarity Coefficient Scores**								**Mean**	**Class-wise DSC**		
	DSC	**Aorta**	**GB**	**KL**	**KR**	**Liver**	**PC**	**SP**	**SM**	**DSC**	**MYO**	**RL**	**LV**
UNet	77.54	85.52	**61.86**	80.57	77.24	94.37	54.72	**87.95**	78.12	88.88	86.89	85.20	94.55
TransUNet	77.48	**87.23**	63.13	81.87	77.02	94.08	55.86	85.08	75.62	89.69	87.42	86.80	94.88
UNet(ours)	**78.36**	**86.42**	61.16	**83.55**	**79.64**	**94.44**	**57.69**	85.67	**78.32**	**89.69**	**87.90**	**86.62**	**94.74**
TransUNet(ours)	**78.74**	85.79	**63.61**	**82.73**	**77.38**	**94.90**	**59.09**	**86.44**	**80.00**	**91.32**	**90.09**	**88.34**	**95.53**
	HD	**Class-wise Hausdorff Distance Scores**								**HD**	**Class-wise HD**		
UNet	38.26	8.06	54.21	**44.52**	75.69	33.67	16.92	**47.81**	25.17	1.98	3.81	1.10	**1.05**
TransUNet	**30.45**	15.65	**38.33**	51.51	48.77	**20.21**	15.05	**38.71**	**15.34**	1.82	3.39	1.06	**1.04**
UNet(ours)	**32.48**	**7.17**	**34.37**	48.99	**64.63**	**22.09**	**11.82**	49.36	**21.42**	**1.54**	**2.49**	**1.07**	1.08
TransUNet(ours)	33.63	**11.32**	44.52	**50.93**	**38.68**	23.88	**13.67**	68.25	17.77	**1.30**	**1.85**	**1.02**	**1.04**

4.2 Robustness Against Inter-class Independence

We comprehensively analyze the effectiveness of our approach in modeling inter-class dependencies. The results shown in Table 2 distinctly showcase the advantages of synergistically employing both spatial and proposed spectral ($\mathcal{L}_{spectral}$) objectives. Our approach proves to be highly effective in mitigating issues arising from the dependencies between different classes in the segmentation process. Furthermore, it exhibits a clear superiority in accurately delineating both larger, more general objects and intricate fine boundaries between objects.

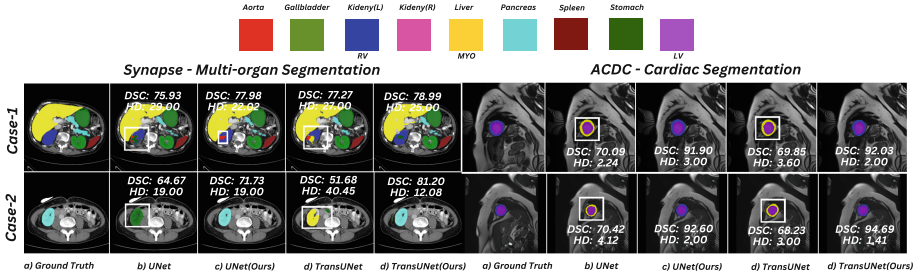


Fig. 7. Segmentation maps on ACDC and Synapse datasets. The segmentation maps are color-coded to represent different anatomical structures. The overlapping white bounding box represents errors made by the respective model. (Color figure online)

Results on Synapse. We observed noteworthy improvements in segmentation performance for both UNet (ours) and TransUNet (ours), compared to UNet and TransUNet. UNet (ours), a CNN architecture, improves DSC by 0.82pp and reduces HD by 5.78pp Interestingly, TransUNet (Ours), a transformer-based architecture, demonstrates a substantial 1.26 pp improvement in DSC, with a surprising increase of 3.18 pp in HD. This suggests that the efficacy of synergizing spatial-spectral representations depends on the specific architecture employed.

The CNNs excel at encoding local information, yet they often struggle to effectively capture global context. In contrast, transformer models are adept at modeling global relationships within data. Our findings reveal that optimizing UNet using $\mathcal{L}_{final}$ led to considerable progress in accurately delineating organs of varying sizes and in capturing intricate fine boundaries. This is due to the spectral regularizer's ability to model contextual long-range dependencies, providing a complementary regularization effect to CNN's strengths. However, TransUNet (ours) tends to overly rely on the global context. While this improved organ delineation, it also led to a limitation in accurately delineating boundaries in multi-organ segmentation scenarios. Figure 7 further highlights that when equipped with the proposed objective, existing networks can effectively model inter-class dependencies and improved delineation of organs and boundaries.

Results on ACDC: The trends are similar to those on the Synapse dataset. UNet(ours) improves DSC by a substantial 0.81 pp and reduces HD by 0.44 pp. Interestingly, TransUNet(ours) improves DSC by a substantial 1.63 pp and reduces HD by 0.52 pp. On ACDC dataset, both UNet (ours) and TransUNet (ours) show improved performance in delineating complex multi-scale contour boundaries. The superior performance of TransUNet (ours) suggests that the multi-scale nature of cardiac structure benefits more from the middle to the global context. Figure 7 highlights that our objective helps existing networks to better model inter-class dependencies and overlapping structures delineation.

Out-of-Distribution Robustness Table 3 shows the results obtained when models pre-trained on ISIC-18 and Kvasir-SEG are tested on the ISIC-17 and PolypGen datasets. For the ISIC-17 dataset, both UNet(ours) and TransUNet(ours) demonstrate substantial improvements in both DSC and IOU. TransUNet(ours) is more sensitive compared to others. Moving to the more challenging PolypGen dataset, which comprises polyp data from 6 different centers, we observe a different trend. Specifically, UNet(ours) demonstrates lower generalization capacities compared to UNet. In contrast, TransUNet(ours) exhibits much stronger generalization capabilities. Quantitatively, we observe a 3.72 pp improvement in DSC, a 4.04 pp improvement in IOU, and a 1.49 pp improvement in sensitivity.

Table 3. OOD testing results: **ISIC-18 → ISIC-17** (pre-trained on ISIC-18 and tested on ISIC-17) and **Kvasir-SEG → PolypGen**.

Method	ISIC-18 → ISIC-17				Kvasir-SEG → PolypGen			
	DSC	IOU	SE	SP	DSC	IOU	SE	SP
UNet	94.01	76.65	80.50	**98.25**	**43.98**	**37.15**	**45.77**	96.24
TransUNet	93.61	76.41	**82.84**	96.85	39.99	32.92	44.14	95.05
UNet(ours)	**94.38**	**77.20**	**82.05**	97.84	40.35	33.81	42.06	**96.27**
TransUNet(ours)	**94.86**	**77.84**	82.62	**97.78**	**43.71**	**36.96**	**45.63**	**96.63**

The difference in performance between UNet(ours) and UNet could be attributed to the fact that the proposed objective $\mathcal{L}_{spectral}$, as discussed earlier, is designed to capture relationships and variations between objects present in the mask. However, this dataset may lack such variations or may not have a sufficient amount of them, leading to the observed performance difference. On the other hand, the improved generalization capabilities of TransUNet (ours) can be attributed to the transformer's ability to capture long-range dependencies, in addition to the contribution from middle-order features from $\mathcal{L}_{spectral}$. Figure 5 provides additional visual evidence of the enhanced capabilities of UNet (ours) and TransUNet (ours) in accurately delineating diverse objects within an out-of-distribution (OOD) setting. These results highlight the versatility of our approach in addressing segmentation challenges across datasets with varying characteristics and complexities. The consistent improvements in performance on both ISIC-17 and PolypGen datasets underscore the generalizability and effectiveness of our proposed method.

4.3 Calibration Analysis

We comprehensively evaluate the effectiveness of our proposed approach in generating confident predictions. Table 4 shows calibration results for both the In-Distribution (IID) and Out-of-Distribution (OOD) settings. For the IID-setting datasets ISIC-18 and Kvasir-SEG, both UNet(ours) and TransUNet(ours) generate confident predictions compared to their respective baselines. However, on the Kviser dataset, UNet(ours) provides comparable results to UNet, whereas TransUNet(ours) reduces ECE and MCE by 1.37pp and 2.59pp, respectively. This suggests that under the IID setting, both models consistently provide confident predictions, but their performance varies based on the dataset characteristics.

Table 4. Calibration performance under IID and OOD setting.

Method	ISIC-18		Kvasir-SEG		ISIC-18 → ISIC-17		Kviser → PolypGen	
	ECE	MCE	ECE	MCE	ECE	MCE	ECE	MCE
UNet	9.13	17.51	8.61	**15.85**	**12.80**	**25.00**	23.67	44.53
TransUNet	9.72	18.60	9.47	17.86	14.04	27.50	27.33	51.71
UNet(ours)	**8.68**	**16.60**	**8.48**	16.04	13.34	26.12	**23.30**	**43.49**
TransUNet(ours)	**9.46**	**18.13**	**8.10**	**15.27**	**13.10**	**25.54**	**21.19**	**39.58**

Under the OOD setting, on the ISIC-17 dataset, UNet(ours) generates less confident predictions than UNet, while TransUNet(ours) again shows improved calibration. TransUNet(ours) reduces ECE and MCE by 0.94pp and 1.96pp, respectively. For the PolypGen dataset, TransUNet generates highly confident predictions, while UNet exhibits slightly more sensitive behavior and demonstrates slightly improved calibration. In summary, our proposed approach consistently yields confident predictions under both IID and OOD settings.

4.4 Robustness Against Noise

MRI and CT scan images are often imperfect due to hardware limitations and patient motion. To test the resilience of our approach, we simulate synthetic Gaussian and Bernoulli noise. Noise levels are set to 0.01, in line with real-world artifacts. As shown in Table 5, UNet (ours) demonstrates improved robustness against noise for the Synapse dataset. However, TransUNet(ours) loses boundary details (higher HD). For the ACDC dataset, both UNet(ours) and TransUNet(ours) demonstrate improved robustness against noise. The pattern remains same for both the noise types. These results strongly emphasize our proposed objective's efficacy in improving models' generalization under noisy conditions.

Table 5. Performance comparison under noise.

	Gaussian				Bernaulli			
Method	Synapse		ACDC		Synapse		ACDC	
	DSC	HD	DSC	HD	DSC	HD	DSC	HD
UNet	70.37	37.82	**73.59**	3.72	76.19	44.93	41.60	7.16
TransUNet	66.59	**30.15**	76.62	3.04	72.02	**42.68**	49.55	5.79
UNet(ours)	**71.78**	**29.51**	73.46	**2.64**	**77.07**	**32.82**	**46.26**	**6.37**
TransUNet(ours)	**70.74**	36.63	**79.18**	**2.61**	**76.49**	49.28	**53.98**	**5.15**

4.5 Interpretability Analysis

Acquired Qualitative Spectral Maps. Figure 8 compares each model's spectral maps for Synapse and ACDC datasets. The proposed UNet(ours) and TransUNet(ours) better preserve low to high frequencies compared to the baselines.

This improved preservation of spectral information contributes to a higher correlation with the ground-truth spectral map. This makes the predictions more interpretable and aligned with the underlying anatomical structures.

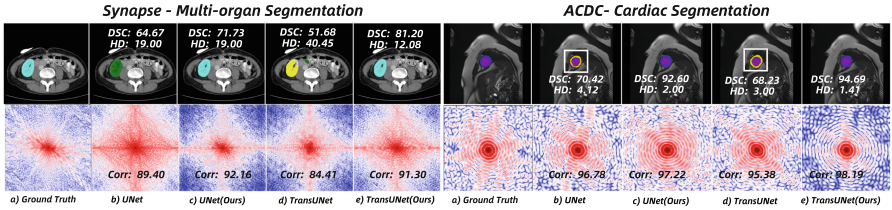


Fig. 8. Spectral maps for Synapse (row 1 and 2) and ACDC (row 3 and 4) datasets (Corr. = correlation).

Acquired Gradient-weighted Class Activation Maps (CAMs): From CAMs provided in Fig. 9, we conclude that: (1) UNet, with its limited receptive field, focuses on local context and overlooks global context, which is crucial for tasks like multi-organ segmentation. Our proposed spectral regularizer enhances UNet's capacity to capture both global contextual relationships while preserving local details across variations. (2) TransUNet tends to emphasize irrelevant regions due to non-contextual long-range dependency modeling. In contrast, our TransUNet(ours) excels at modeling contextual long-range dependencies and middle-order features, thereby attending to both local and global contexts. (3) Our method notably excels in modeling intra-class variations across sequences, surpassing baselines. (4) Unet(ours) offers higher interpretability compared to TransUNet and remains competitive with TransUNet(ours).

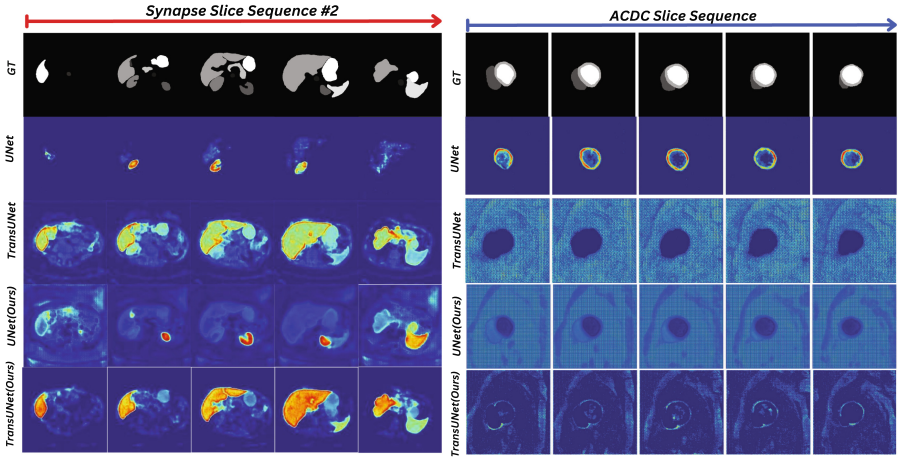


Fig. 9. Gradient-weighted class activation maps

4.6 Sensitivity Analysis

Table 6. Sensitivity Analysis

Backbone	ACDC		Kviser-SEG	
	DSC	HD	DSC	IOU
$\lambda = 0.1$	88.96	1.88	88.35	76.84
$\lambda = 0.2$	**89.69**	**1.54**	**89.40**	77.03
$\lambda = 0.3$	89.33	1.74	88.26	76.84
$\lambda = 0.5$	87.54	2.10	87.34	**77.34**
$\lambda = 0.9$	79.88	3.77	84.70	73.46

In Sect. 2.3 of the manuscript, we introduce the λ hyperparameter, and in Table 6, we present a comprehensive analysis of its impact on the ACDC and Kviser-SEG datasets. For the ACDC dataset, our analysis reveals that setting λ to 0.2 yields the most favorable results in terms of Dice Similarity Coefficient (DSC) and Hausdorff Distance (HD). On the other hand, when considering the Kviser-SEG dataset, we found that a value of $\lambda = 0.2$ leads to the highest DSC, whereas $\lambda = 0.5$ produces the best Intersection over Union (IOU). Notably, as we increase the value of λ beyond these optimal settings, we observe a noticeable degradation in segmentation performance for both datasets. This observation underscores the significance of spatial representation in segmentation tasks, indicating that spatial features offer superior representations. However, it is worth emphasizing that a judiciously crafted weighting scheme that synergizes spatial and spectral information can potentially enhance domain generalization.

5 Conclusion, Limitations and Future Work

In this study, we introduce a novel spectral objective, the spectral correlation coefficient, in synergy with a spatial objective, effectively enhancing domain generalization in medical image segmentation. This approach seamlessly integrates with existing encoder-decoder architectures. When combined with TransUNet, it achieves remarkable performance and outperforms state-of-the-art methods across diverse medical segmentation tasks. Our method exhibits interpretability and resilience to noisy data while generating confident predictions. Future work will concentrate on minimizing false negatives, especially in noisy environments. One intriguing avenue for future research involves integrating our proposed method into established semi-supervised or knowledge distillation-based [11] approaches to increase efficiency in terms of annotation and computation. Additionally, there is potential for extending the scope of our method beyond medical tasks, conducting performance analyses in diverse application domains.

Acknowledgement. This study was supported by NIH grants R01-CA246704, R01-CA240639, U01 DK127384-02S1, and U01-CA268808. Sparsh is supported by the SERB project CRG/2022/003821. IIT Roorkee provided support for the computing system used for this research under the grant FIG-100874. Vandan, Sparsh, and Ulas are the corresponding authors.

References

1. Multi-atlas abdomen labeling challenge. synapse multi-organ segmentation dataset (2015). https://www.synapse.org/#!Synapse:syn3193805/wiki/217789
2. ACDC (automated cardiac diagnosis challenge) (2017). https://www.creatis.insa-lyon.fr/Challenge/acdc
3. Ali, S., et al.: A multi-centre polyp detection and segmentation dataset for generalisability assessment. Sci. Data **10**(1), 75 (2023)
4. Azad, R., Asadi-Aghbolaghi, M., Fathy, M., Escalera, S.: Bi-directional convlstm u-net with densley connected convolutions. In: Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops (2019)
5. Bao, H., Dong, L., Piao, S., Wei, F.: Beit: BERT pre-training of image transformers. arXiv preprint arXiv:2106.08254 (2021)
6. Basak, H., Kundu, R., Sarkar, R.: MFSNet: a multi focus segmentation network for skin lesion segmentation. Pattern Recogn. **128**, 108673 (2022)
7. Chen, J., et al.: Transunet: transformers make strong encoders for medical image segmentation. arXiv preprint arXiv:2102.04306 (2021)
8. Cheng, J.: Brain tumor dataset (2017). https://doi.org/10.6084/m9.figshare.1512427.v5
9. 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 (2018)
10. Das, A., et al.: Pam-unet: shifting attention on region of interest in medical images. arXiv preprint arXiv:2405.01503 (2024)
11. Gorade, V., Mittal, S., Jha, D., Bagci, U.: Rethinking intermediate layers design in knowledge distillation for kidney and liver tumor segmentation
12. Gorade, V., Mittal, S., Jha, D., Bagci, U.: Synergynet: bridging the gap between discrete and continuous representations for precise medical image segmentation. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pp. 7768–7777 (2024)
13. Gorade, V., Mittal, S., Singhal, R.: PACL: patient-aware contrastive learning through metadata refinement for generalized early disease diagnosis. Comput. Biol. Med. **167**, 107569 (2023)
14. Gorade, V., Singh, A., Mishra, D.: Large scale time-series representation learning via simultaneous low-and high-frequency feature bootstrapping. IEEE Trans. Neural Netw. Learn. Syst. (2023)
15. Gorade, V., et al.: Towards synergistic deep learning models for volumetric cirrhotic liver segmentation in MRIs. arXiv preprint arXiv:2408.04491 (2024)
16. Huang, H., et al.: Unet 3+: a full-scale connected unet for medical image segmentation. In: 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), ICASSP 2020, pp. 1055–1059. IEEE (2020)
17. Jha, D., et al.: Kvasir-seg: a segmented polyp dataset. In: Proceedings of the 26th International Conference on MultiMedia Modeling, pp. 451–462 (2020)

18. Kawaguchi, K., Kaelbling, L.P., Bengio, Y.: Generalization in deep learning. arXiv preprint arXiv:1710.05468, **1**(8) (2017)
19. Kumar, N., et al.: A multi-organ nucleus segmentation challenge. IEEE Trans. Med. Imaging **39**(5), 1380–1391 (2019)
20. Li, P., Zhou, R., He, J., Zhao, S., Tian, Y.: A global-frequency-domain network for medical image segmentation. Comput. Biol. Med. 107290 (2023)
21. Mendonça, T., Ferreira, P.M., Marques, J.S., Marcal, A.R., Rozeira, J.: Ph 2-a dermoscopic image database for research and benchmarking. In: 2013 35th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), pp. 5437–5440 (2013)
22. Morrison, K., Gilby, B., Lipchak, C., Mattioli, A., Kovashka, A.: Exploring corruption robustness: inductive biases in vision transformers and MLP-mixers. arXiv preprint arXiv:2106.13122 (2021)
23. Neyshabur, B., Bhojanapalli, S., McAllester, D., Srebro, N.: Exploring generalization in deep learning. In: Advances in Neural Information Processing Systems, vol. 30 (2017)
24. Park, N., Kim, S.: How do vision transformers work? arXiv preprint arXiv:2202.06709 (2022)
25. Raghu, M., Unterthiner, T., Kornblith, S., Zhang, C., Dosovitskiy, A.: Do vision transformers see like convolutional neural networks? In: Advances in Neural Information Processing Systems, vol. 34, pp. 12116–12128 (2021)
26. Wang, Z., Wu, L.: Theoretical analysis of inductive biases in deep convolutional networks. arXiv preprint arXiv:2305.08404 (2023)
27. Wei, C., Fan, H., Xie, S., Wu, C.Y., Yuille, A., Feichtenhofer, C.: Masked feature prediction for self-supervised visual pre-training. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 14668–14678 (2022)
28. Yang, C., Guo, X., Chen, Z., Yuan, Y.: Source free domain adaptation for medical image segmentation with Fourier style mining. Med. Image Anal. **79**, 102457 (2022)
29. Yang, Y., Soatto, S.: FDA: Fourier domain adaptation for semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 4085–4095 (2020)

Leveraging Point Annotations in Segmentation Learning with Boundary Loss

Eva Breznik[1,2(✉)], Hoel Kervadec[5], Filip Malmberg[1], Joel Kullberg[3,4], Håkan Ahlström[4], Marleen de Bruijne[5,6], and Robin Strand[1,4]

[1] Department of Information Technology, Uppsala University, Uppsala, Sweden
{eva.breznik,robin.strand}@it.uu.se
[2] Department of Biomedical Engineering and Health Systems, Royal Institute of Technology, Stockholm, Sweden
[3] Antaros Medical, Mölndal, Sweden
[4] Department of Surgical Sciences, Uppsala University, Uppsala, Sweden
[5] Department of Radiology and Nuclear Medicine, Erasmus MC, Rotterdam, The Netherlands
[6] Department of Computer Science, University of Copenhagen, Copenhagen, Denmark

Abstract. This paper investigates the combination of intensity-based distance maps with boundary loss for point-supervised semantic segmentation. By design, the boundary loss imposes a stronger penalty on the errors the farther away from the object boundary they occur. Hence it is inappropriate for cases of weak supervision where the ground truth label is much smaller than the actual object and a certain amount of false positives (w.r.t. the weak ground truth) is actually desirable. Using intensity-aware distances instead may alleviate this drawback, allowing for a certain amount of false positives with similar intensities without a significant increase to the training loss. This formulation is potentially more attractive than existing CRF-based regularizers, due to its simplicity and computational efficiency. We perform experiments on two multi-class datasets; ACDC (heart segmentation) and POEM (whole-body abdominal organ segmentation). Results are encouraging and show that this supervision strategy has great potential. On ACDC it outperforms the CRF-loss based approach, and on POEM data it performs on par with it. The code is made openly available.

Keywords: Segmentation · Point supervision · Boundary loss · Minimum barrier distance

1 Introduction

Convolutional neural networks (CNNs) are now the method of choice for various image processing tasks including segmentation. However, they typically require a

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/978-3-031-78201-5_13.

A. Antonacopoulos et al. (Eds.): ICPR 2024, LNCS 15313, pp. 194–210, 2025.
https://doi.org/10.1007/978-3-031-78201-5_13

large amount of annotated ground truth data for training. In the medical domain in particular such annotations require expert knowledge and are very costly to obtain. This increased research interest in weakly supervised training, aiming to utilize approximate labels that are cheaper and faster to produce.

Weak supervision comes in many different flavors, for example image-level labels [8,9,32], bounding boxes [7,15,26,32], scribbles [12,14,18,28] or point annotations [3,5,16,25], to name a few. Existing weakly supervised methods for image segmentation typically either introduce additional segmentation constraints or regularization [3,14,15,35], or generate pseudo-labels as an exact ground truth substitute to use in full supervision [12,16,18,20,33]. But outside priors or additional information is generally required to formulate useful constraints and training with inaccurate labels can propagate errors, causing instabilities.

With the aim to improve the accuracy of segmentation methods, distance maps have been employed in a multitude of ways [21]. An example of a very direct inclusion of a distance map to guide the CNN training is boundary loss [13], introduced in context of fully supervised segmentation. It uses a Euclidean distance transform of the ground truth to minimize the distance between the ground truth and the predicted segmentation boundaries at training time. This proved very effective while remaining computationally lightweight and compatible with different network architectures, optimization strategies and other losses.

In this paper, we investigate the use of boundary loss in a weakly supervised setting and propose a way to make it directly compatible with training on point annotations without requiring architectural changes or modifications to the training procedure. We propose to replace the Euclidean distance map used in the original boundary loss paper with intensity-aware ones, taking pixel intensities into account when computing the distance from the point annotations. In applications where the object to segment has a fairly homogeneous intensity and a decent contrast around the object boundary, this can provide the network with a better notion of the region extent and shape, addressing the incompatibility between boundary loss and weak labels. It allows end-to-end training without explicitly formulated priors, additional data-dependent information or pseudo-label creation.

Our proposed approach is evaluated using various intensity-aware distances, on two multi-class segmentation tasks with artificially created point annotations. We show that training with the combination of cross entropy and boundary loss is not only compatible with point-level supervision but reaches competitive results compared to a CRF-loss based training and even compared to full supervision.

2 Related Works

2.1 Point and Scribble Supervision

Xu et al. [32] present a unified approach for various types of weak supervision like bounding boxes, image tags and partial labels. Their approach is based on first oversegmenting the image to superpixels, then using max-margin clustering with weak annotations as constraints.

ScribbleSup [18] makes use of a superpixel-based graphical model for annotation propagation jointly with a CNN based segmentation model, optimizing them alternately. While the method was developed specifically for training with scribbles, they report good results even on point annotations. Using point-level supervision, [3] introduce a special loss based on both global image-level labels and weighted supervision on annotated pixels. The loss is further supplemented by an objectness prior, which requires independent supervised pretraining.

Conditional random field (CRF) based losses [29,34] implement the seminal work of [17], and using such losses to add regularization has been shown to achieve good results under weak supervision. In [34], dense CRFs are reformulated as RNNs, to achieve an end-to-end trainable segmentation system. Tang et al. [28] propose a kernel cut loss combining CRF and normalized cut terms and apply it together with cross-entropy loss for semantic segmentation tasks with scribble annotations.

While many works (e.g. [12,18]) use CRF as a postprocessing step to further improve the segmentation results, it was shown that incorporating it as loss can still be advantageous [29]. Also based on point annotations and CRF loss, [25] create two coarse segmentations to be used in training: a Voronoi diagram based one as a lower bound (undersegmentation) and a clustering based one as an upper bound (oversegmentation). They show that training with both segmentations jointly, with the addition of a CRF loss, performs well on nuclei segmentation in histopathology images. All mentioned CRF-based methods however significantly slow down the training time per iteration.

In [5], superpixels are created and labeled according to the point annotations, using learned hierarchical features and superpixel affinity to propagate the labels into the initially un-annotated superpixels. Liu et al. [20] focus on nuclei (single class) instance segmentation, proposing a two-step learning scheme combining a foreground proposal model with an instance separation model, both trained with pseudo-labels. Even Yao et al. [33] rely on pseudo-label supervision for nuclei segmentation, but add a so-called anchor quality loss to supress spurious responses far away from the point labels. This implicitly assumes the point annotations are close to object centroids, which is conceptually similar to using boundary loss with Eclidean distance on point annotations. In [16], a small set of fully annotated data is used in stacked object proposal and refinment models of a teacher network to generate pseudo-labels. The final student network is trained with both pseudo- and full labels.

2.2 Non-euclidean Distance Maps in Segmentation Tasks

Among works employing distance maps to boost segmentation, [6] apply a Geodesic distance-based filtering operator to produce a set of smooth segmentation proposals, a viable subset of which is then searched for the best, energy minimizing labelling. In [2] and [10] the authors make use of Geodesic maps for scribble-based segmentation, but in an interactive setting. While [10] introduces a geodesic star convexity shape constraint computed on (intensity/colour-based) likelihood maps, the geodesic matting framework from [2] calculates the Geodesic

map on the space of the class probability densities and models class distributions based on the user scribble statistics. Also in a purely interactive setting, [31] use the Geodesic maps from user-provided scribbles as additional input channels during CNN refinement. In [22], use of Geodesic map priors is proposed to improve robustness when training with noisy labels. Their method consists of first training an autoencoder to regress ground truth annotations based on the Geodesic maps, then using the mean square error between the encodings of segmentor network probabilities and distance maps as a part of the loss during segmentor network training. Their approach requires additional training time and while they work with noisy labels, they do not cover severe label degradation such as scribble or point annotations. In [23] the authors propose a vectorial Dahu pseudo-distance, based on the Minimum barrier distance. While developed and evaluated mainly for saliency detection, it is successfully used also in white-matter segmentation with simple seeding and thresholding.

3 Method

3.1 The Boundary Loss

Boundary loss from [13] has so far been successfully applied in CNN training for fully supervised segmentation tasks. It is calculated by computing a signed distance map of the ground truth boundary mask and multiplying it with the network output.

Formally, we have $\Omega \subset \mathbb{R}^D$ as D-dimensional image space, with $X : \Omega \rightarrow \mathbb{R}^M$ an image with M modalities, and $Y : \Omega \rightarrow \mathcal{K}$ its corresponding multi-class ground truth, with $\mathcal{K} = \{0 : \texttt{background}, 1 : \texttt{first class}, ..., K : \texttt{last class}\}$. For simplicity, we will denote $Y^{(k)} := \{i \in \Omega | Y(i) = k\}$ the subset of Ω containing all the voxels belonging to the class k. It follows that $\cup\{Y^{(k)}\}_{k \in \mathcal{K}} = \Omega$, and $Y^{(i)} \cap Y^{(j)} = \emptyset$ for any pair of classes $i \neq j; i, j \in \mathcal{K}$ (i.e. they do not overlap).

In the original work on boundary loss the signed distance map for each class k is computed as follows:

$$\forall i \in \Omega : \phi_Y^{(k)}(i) = \begin{cases} -D_{\text{euc}}^{(k)}(i) & \text{if } i \in Y^{(k)}; \\ D_{\text{euc}}^{(k)}(i) & \text{otherwise,} \end{cases} \tag{1}$$

with $D_{\text{euc}}^{(k)}(\cdot) : \Omega \rightarrow \mathbb{R}_+$ denoting the Euclidean distance map from the boundary of the ground truth annotation for class k. Strictly at the annotation boundary $\partial Y^{(k)}$, the value of the distance map is 0 (with the boundary $\partial Y^{(k)}$ here denoting all elements of $Y^{(k)}$ that have a neighbor outside this set).

Then, this distance map is used as-is inside the boundary loss:

$$\mathcal{L}_B(s_{\boldsymbol{\theta}}, Y) = \sum_{k \in \mathcal{K}} \sum_{i \in \Omega} s_{\boldsymbol{\theta}}^{(i,k)} \phi_Y^{(k)}(i), \tag{2}$$

where $s_{\boldsymbol{\theta}}$ represents the probabilities predicted by the network.

3.2 From Full- to Point Annotation

Generally, given an exact, full annotation of an object, such a Euclidean signed distance map encodes the object shape. But when using a weak ground truth $\widetilde{Y} : \Omega \to K$, where $\widetilde{Y}^{(k)} \subset Y^{(k)}$ and $\cup\{\widetilde{Y}^{(k)}\}_{k\in\mathcal{K}} \neq \Omega$, the distance simply grows radially regardless of the actual shape of the object, thus making little sense from an information point of view. Under the assumption of intra-object homogeneity and inter-object intensity contrast, this problem can be circumvented to a degree by using a distance function that takes also intensity values into account. An example of a commonly used distance measure with an intensity component is the gray scale Geodesic distance [30]. Let $\pi_{x,y} = \langle x = p_0, \ldots, p_i, p_{i+1}, \ldots, p_n = y\rangle$ denote a path between $x, y \in \Omega$, with p_i and p_{i+1} being neighbors under a chosen adjacency relation. Reusing the notation from before, a Geodesic distance map from the boundary of the ground truth class k, $D^{(k)}_{\text{geo}}(\cdot) : \Omega \to \mathbb{R}_+$, can be defined as

$$D^{(k)}_{\text{geo}}(\cdot) = \min_{\pi\in\Pi_k(\cdot)} \sum_{p_n,p_{n+1}\in\pi} \sqrt{(1-\lambda)\Delta I^2_{n,n+1} + \lambda c^2_{n,n+1}} \tag{3}$$

where $\Pi_k(\cdot) := \cup_{j\in\partial Y^{(k)}} \pi_{\cdot,j}$ and $\Delta I_{n,n+1} = I_n - I_{n+1}$ is the intensity difference between $p_n, p_{n+1} \in \Omega$. The value of $c_{n,n+1}$ is the distance between p_n and p_{n+1} in space, and depends on the type of adjacency between them (in 3D for example, it is equal to 1, $\sqrt{2}$ or $\sqrt{3}$ if p_n and p_{n+1} share a face, an edge or a vertex respectively). The parameter $\lambda \in [0, 1]$ allows for balancing the contributions of intensities and spatial proximity.

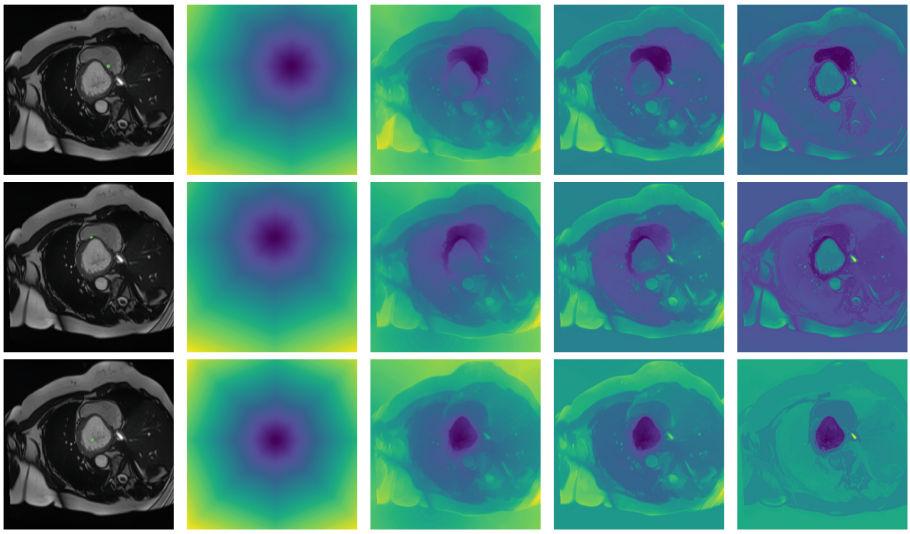
In practice, the Geodesic distance is often implemented using a weighted L_1 distance instead. That means changing the expression under the sum in Eq. (3) to $(1-\lambda)|\Delta I_{n,n+1}| + \lambda c_{n,n+1}$. This definition is adopted also throughout this paper, as it is easier and faster to compute and locally approximates the original one up to the next greater integer number (see e.g. the implementation of [1]).

While setting the λ parameter to 1 actually results in a taxicab distance, which can be seen as a discrete approximation of a Euclidean distance, setting $\lambda = 0$ focuses purely on image intensities. From here on we shall call the latter simply Intensity distance, D_{int}, and use the term Geodesic distance, D_{geo}, to denote the setting of $\lambda = 0.5$.

Entirely on the other side of the spectrum from the Euclidean are the fully intensity-based distances, such as the Minimum barrier distance (MBD) from [27]. It is calculated exclusively on the image intensity space and effectively independent of the path length in space. Given the definition of $\Pi_k(\cdot)$ above, the MBD map from the boundary of ground truth class k, $D^{(k)}_{\text{mbd}}(\cdot) : \Omega \to \mathbb{R}_+$, can be defined as:

$$D^{(k)}_{\text{mbd}}(\cdot) = \min_{\pi\in\Pi_k(\cdot)} \left(\max_{p_i\in\pi} I_{p_i} - \min_{p_j\in\pi} I_{p_j} \right). \tag{4}$$

MBD and Intensity distance are only pseudo-distances, as the reflexivity property may be violated (i.e. it may be that $D^{(k)}_{\text{mbd/int}}(x) = 0$ even when $x \notin$



(a) Point labels (b) D_{euc} maps (c) D_{geo} maps (d) D_{int} maps (e) D_{mbd} maps

Fig. 1. Comparison of different distance maps on an example slice from the ACDC dataset [4], computed from the point labels shown in the first column. The three rows show the ground truth and corresponding distance maps for the right ventricle, myocardium and left ventricle respectively.

$Y^{(k)}$). As opposed to full supervision, this can be a very desirable property in the case of weak labels.

Both the Intensity and the Minimum barrier distance are defined exclusively on the image intensity space. However, from the examples of D_{int} distance map in Fig. 1d, we can notice that the values still increase somewhat radially from the annotation. This behaviour is similar to the one of the Geodesic distance in Fig. 1c (which actually includes the spatial proximity in its definition), and is due to the summing operator in the general Geodesic distance definition in Equation 3. While the intensities of two neighboring pixels on a path may be the same, that will rarely be the case in real life, noise riddled images. This makes the Intensity distance function approximately monotonically increasing with increasing length of the path (in space), even in fairly homogeneous regions, which is a potential drawback under point supervision.

As opposed to the Euclidean distance, all intensity-aware distances (geodesic, intensity and MBD) are able to encode contrast sensitivity end preserve object structure by harnessing intensity information. But MBD is the only distance entirely disregarding spatial information, resulting in a less pronounced and smooth increase in the values outward from the source point. In practice, using such maps for network training means a lower penalty for false positives that occur farther from the point annotation but are close to it in intensity, which is desirable under weak supervision.

If we were to use the boundary loss on the point labels alone, very few pixels would be *positively* supervised: their probability is pushed *up* only when the distance map is negative, i.e. on the exact dot annotation. We solve this minor issue by combining it with a partial cross-entropy, $\mathcal{L}_{\widetilde{\mathrm{CE}}}$, which results in the following model to optimize:

$$\min_{\boldsymbol{\theta}} \sum_{k \in \mathcal{K}} \sum_{i \in Y^{(k)}} -\log(s_{\boldsymbol{\theta}}^{(i,k)}) + \alpha \mathcal{L}_B, \tag{5}$$

with $\alpha \in \mathbb{R}$ balancing the two losses. To differentiate between the boundary loss computation with different distances, we use $\mathcal{L}_{\mathrm{B}}^{d}$ to denote the boundary loss computed with the distance metric $d \in \{\text{euc, int, geo, mbd}\}$.

The different distance functions have different advantages and drawbacks, and when using them on real-life medical data, there are additional considerations that need to be taken into account when choosing or designing the distance metric for a particular use case. For example image dimensions and field of view, sampling, number of modalities/channels, and distance range and stability. They are described in more detail in the supplementary material.

4 Experiments

4.1 Datasets

ACDC [4]**.** The Automated Cardiac Diagnosis Challenge (ACDC) is a public benchmark multi-class heart segmentation dataset. It contains cine-MR images of 150 patients of which 100 are available for training, covering healthy scans and four types of pathologies in equal amounts, with annotations for the right ventricle (RV), myocardium (Myo) and left ventricle (LV) heart structures. We split the training set randomly, using 65 subjects for training, 10 for validation and 25 as a hold-out test set. Due to the large and varying interslice gap, we train with (and compute distance maps on) 2D slices.

We normalize the volumes and resize the slices to 256×256 pixels. As the official dataset comes with full annotations, a synthetic point ground truth is created randomly for every slice and every foreground object present in it. For more details on this process see the Supplementary A.

POEM. The Prospective investigation of Obesity, ENergy production and Meta-bolism (POEM) is a local (not currently publicly available; PI: L. Lind, see [19] for details) cohort of whole-body fat/water separated MR images. Full annotations of the liver, kidneys, bladder, pancreas and spleen are available for 50 subjects, providing a challenging segmentation dataset with heavily imbalanced classes of varying shapes. The resolution is anisotropic, with reconstructed voxel size of $2.07 \times 2.07 \times 8.0\,\mathrm{mm}^3$ in left-right, anterior-posterior and foot-head directions, respectively. For additional technical details regarding the acquisition and image specifications see [19].

We split the dataset randomly, with 35, 5 and 10 subjects for training, validation and testing respectively. The images contain two channels, one for water

and one for fat content. For training, we normalize the volumes (per channel) and use 2D slices in the coronal plane, sized 256×256. The weak annotations are created synthetically, following the same procedure as for the ACDC dataset.

4.2 Distance Map Computation

For computing the intensity-aware distance maps, the image intensities are first scaled to 0–255 to ensure that the spatial distances and intensity differences between voxels are comparable in 2D. The distance maps are then computed prior to training, using full connectivity. On POEM data, they are computed on the fat content channel only. For the approximate Euclidean, Geodesic and Intensity distance, we use the FastGeodis implementation of [1]. For the Minimum barrier distance, we use our own implementation[1]. See Supplementary B for more details on distance maps.

For ACDC, we compute the maps on 2D slices. In the case of POEM, however, the majority of 2D slices contain only background, and even those slices that contain foreground never contain all of the classes. As a lack of a class in an image results in a zero distance map (by implementation design), that particular class can be arbitrarily segmented without an increase in the loss. To circumvent this issue, we use a simple map of ones for every absent class, assuring some minimal amount of penalty. In addition, considering that the POEM data is highly anisotropic and its coronal slices may contain unconnected regions belonging to the same class, we run a separate set of experiments with distance maps calculated in 3D volumes, using all slice-wise point annotations. In 3D, since all classes are present in every volume, the problem of zero distance maps and no supervision is avoided entirely.

4.3 Baselines and Settings

We use cross-entropy with full supervision $\mathcal{L}_{\text{CE}}$ as an upper bound. As a lower bound, we train with partial cross-entropy $\mathcal{L}_{\widetilde{\text{CE}}}$ on the point annotations.

In addition, we compare the results to a state-of-the-art method for weak supervision, using a combination of the partial cross entropy and a CRF-loss from [29]. We choose this work in particular because a number of successful works in various weakly supervised segmentation tasks either build upon it (e.g. [25,28]) or use CRFs for postprocessing (e.g. [12,18]). In addition, it is the only available method that requires only a change of loss in a full supervision setting to allow for point supervision.

For both datasets, we use the lightweight E-Net [24] trained with Adam optimizer. In addition, for the ACDC dataset we also use a U-Net configuration as automatically determined by nnU-Net [11]. Both networks are trained with a combination of the partial cross entropy and boundary loss, where both losses are calculated only on the foreground classes and their contributions are weighted by $\alpha = 1$ in E-Net and $\alpha = 0.7$ in nnU-Net. With nnU-Net, we omit deep supervision

[1] Base code available at https://github.com/FilipMalmberg/DistanceTransforms.

and compensate this with increasing the initial learning rate by a tenfold. For the competing method of [29], the parameters (confirmed by a limited grid search) are $w = 2e - 9$, $\sigma_{\text{rgb}} = 15$, $\sigma_{\text{xy}} = 100$, and the scale factor is set to 0.5. All the experiments are run on a single NVIDIA GeForce GTX 3080 Ti using cuDNN v11.8. The code is available at https://github.com/EvaBr/geodesic_bl.

During training, we monitor the batch Dice score on a fully annotated validation set. The model that performs best according to this measure is used for subsequent evaluations on the test dataset. Each experiment is run 3 times for increased repeatability, and evaluation results are averaged over runs.

4.4 Evaluation Metrics

We evaluate the performance of the methods through two standard segmentation metrics; Dice score (DSC) and 95th percentile Hausdorff distance (HD95). While training is performed on 2D slices, the evaluation metrics are reported on full 3D scans, for each foreground class separately as well as averaged over all classes.

DSC. The Dice similarity score measures the overlap between the ground truth volume G and the output segmentation volume S, and is defined as $\frac{2|G \cap S|}{|G|+|S|}$, where $|\cdot|$ denotes the cardinality (in this case the nonzero element count).

HD95. The Hausdorff distance is a dissimilarity measure, representing the distance between the surfaces of G and S. As it is sensitive to outliers, we use the 95th percentile instead of the maximum for computing the directed distances.

5 Results and Discussion

5.1 Segmentation of Cardiac Structures

The average 3D Dice scores and HD95 values on the ACDC test set are given in Tables 1 and 2 for E-Net and nnU-Net respectively. For the distribution boxplots see Supplementary D. We see that, in terms of DSC, the proposed strategy of

Table 1. Mean ↑DSC and ↓HD95 values over five independent runs with E-Net, calculated on 3D volumes of the ACDC test set. Labels RV, Myo and LV represent the right ventricle, myocardium and left ventricle classes respectively. Boxplots for one run showing the distributions over subjects are available in the supplementary.

Method	RV	Myo	LV	All
$\mathcal{L}_{\text{CE}}$ (fully supervised)	↑ 0.7986 ↓ 2.911	↑ 0.8111 ↓ 1.336	↑ 0.8923 ↓ 2.774	↑ 0.8748 ↓ 1.754
$\mathcal{L}_{\widetilde{\text{CE}}}$ (point annotations)	↑ 0.0991 ↓ 91.645	↑ 0.0689 ↓ 82.347	↑ 0.2060 ↓ 88.119	↑ 0.2137 ↓ 65.528
w/$\mathcal{L}_{\text{B}}^{euc}$	↑ 0.6214 ↓ 6.611	↑ 0.6709 ↓ 3.976	↑ 0.8112 ↓ 5.69	↑ 0.7742 ↓ 4.069
w/$\mathcal{L}_{\text{B}}^{geo}$	↑ 0.637 ↓ 8.106	↑ 0.679 ↓ 5.186	↑ 0.82 ↓ 5.529	↑ 0.782 ↓ 4.705
w/$\mathcal{L}_{\text{B}}^{int}$	↑ 0.639 ↓ 9.991	↑ 0.6729 ↓ 5.435	↑ 0.829 ↓ 4.282	↑ 0.7834 ↓ 4.926
w/$\mathcal{L}_{\text{B}}^{mbd}$	↑ 0.6553 ↓ 8.968	↑ 0.6956 ↓ 5.943	↑ 0.8297 ↓ 6.179	↑ 0.7936 ↓ 5.272
w/CRF-loss [29]	↑ 0.2660 ↓ 63.467	↑ 0.5385 ↓ 27.492	↑ 0.8189 ↓ 16.165	↑ 0.4558 ↓ 26.781

Table 2. Mean ↑DSC and ↓HD95 values over five independent runs with nnU-Net, calculated on 3D volumes of the ACDC test set. Labels RV, Myo and LV represent the right ventricle, myocardium and left ventricle classes respectively. Boxplots for one run showing the distributions over subjects are available in the supplementary.

Method	RV	Myo	LV	All
$\mathcal{L}_{\text{CE}}$ (fully supervised)	↑ 0.850 ↓ 1.087	↑ 0.866 ↓ 0.641	↑ 0.931 ↓ 0.652	↑ 0.911 ↓ 0.595
$\mathcal{L}_{\widetilde{\text{CE}}}$ (point annotations)	↑ 0.029 ↓ 69.171	↑ 0.091 ↓ 56.383	↑ 0.281 ↓ 55.842	↑ 0.100 ↓ 45.349
w/$\mathcal{L}_{\text{B}}^{euc}$	↑ 0.003 ↓ 79.127	↑ 0.129 ↓ 44.884	↑ 0.243 ↓ 53.453	↑ 0.338 ↓ 44.366
w/$\mathcal{L}_{\text{B}}^{geo}$	↑ 0.492 ↓ 13.313	↑ 0.560 ↓ 6.881	↑ 0.707 ↓ 6.923	↑ 0.679 ↓ 6.779
w/$\mathcal{L}_{\text{B}}^{int}$	↑ 0.365 ↓ 8.672	↑ 0.441 ↓ 4.267	↑ 0.509 ↓ 5.843	↑ 0.575 ↓ 4.695
w/$\mathcal{L}_{\text{B}}^{mbd}$	↑ 0.479 ↓ 14.038	↑ 0.536 ↓ 8.911	↑ 0.712 ↓ 8.126	↑ 0.687 ↓ 7.844
w/CRF-loss [29]	↑ 0.023 ↓ 79.106	↑ 0.497 ↓ 10.917	↑ 0.680 ↓ 12.268	↑ 0.300 ↓ 25.573

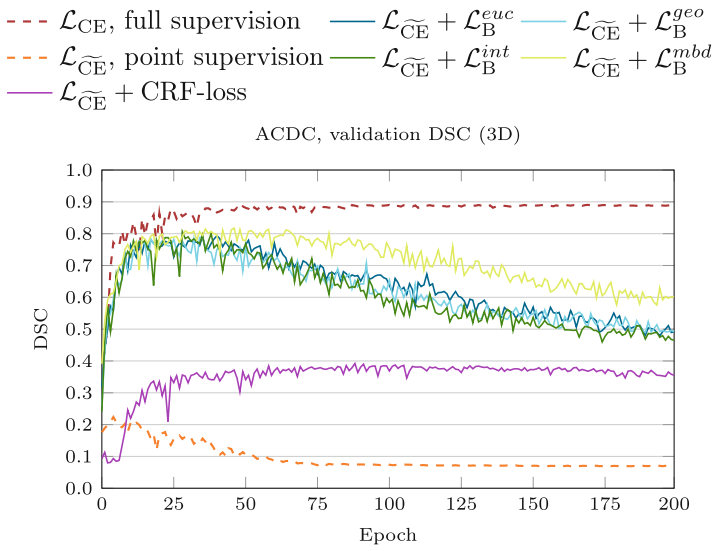


Fig. 2. Curve evolution of the average (over foreground classes) validation (3D) Dice scores during training (on 2D slices) with E-Net, for the ACDC dataset.

using intensity-aware MBD distance within boundary loss performs better than simply using the Euclidean distance, and better than using CRF-loss. The CRF-loss results are significantly worse in both metrics. Figure 4 shows qualitative results on two randomly chosen test slices, confirming that training with $\mathcal{L}_B^{mbd}$ follows the image gradients and recovers the underlying shape better than $\mathcal{L}_B^{euc}$. The CRF-loss recover the shape of the myocardium and left ventricle to some extent, but fails entirely on the right ventricle. In Figs. 2 and 3 we show the 3D DSC validation curve evolution for a single run. The CRF-loss seems to have converged to a low DSC value, regardless of the architecture used. With E-Net, all settings combining CE and boundary loss reach values close to the

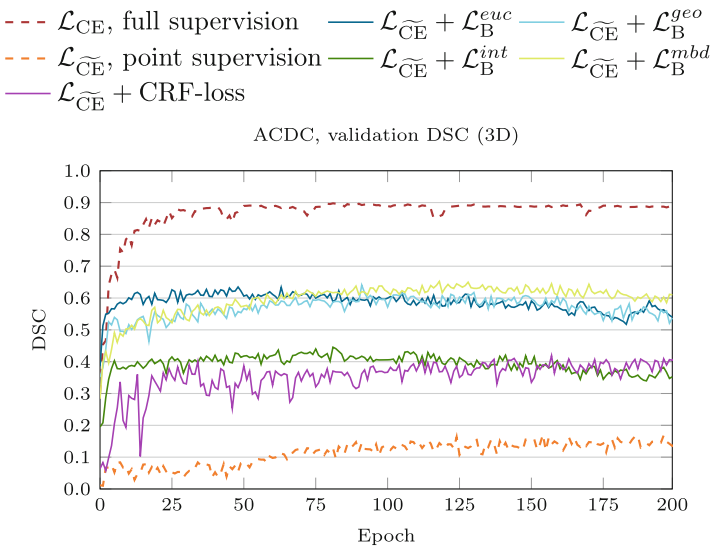


Fig. 3. Curve evolution of the average (over foreground classes) validation (3D) Dice scores during training (on 2D slices) with nnUNet, for the ACDC dataset.

full supervision in the beginning of the training and then slowly collapse towards to the point annotations. The MBD version stands out, degrading slower, thus providing a wider range of potentially good models for evaluation. When using nnU-Net on the other hand, the final performance of the MBD version of the loss is slightly lower, however it is also more stable in terms of degrading slower. CRF is unable to compete with it even here.

5.2 Abdominal Organ Segmentation

Using 2D Distance Maps. Table 3 shows the average DSC and HD95 results (both using 2D and 3D computed distance maps) for the task of abdominal organ segmentation in POEM data, using E-Net (for boxplots see Supplementary D). We see that training with distances calculated on 2D slices $\mathcal{L}_B^{euc}$ and $\mathcal{L}_B^{mbd}$ perform comparably, while $\mathcal{L}_B^{int}$ and $\mathcal{L}_B^{geo}$ lag behind in both DSC and HD95 metric.

On this dataset, the CRF-loss is able to compete with the boundary loss-based training strategies, even outperforming them on most classes. The reason behind its increased performance on the POEM dataset may be due to a larger number of classes, and thus inherently more supervision. Most notably, all models trained with boundary loss appear to have a hard time segmenting the liver. We hypothesize this may be due to extremely severe class imbalance, as the liver covers a very large area compared to the rest of the classes. It is thus also more strongly affected by undersegmentations. According to the validation curves in Fig. 5, training on this dataset is less stable and slower than on ACDC for all

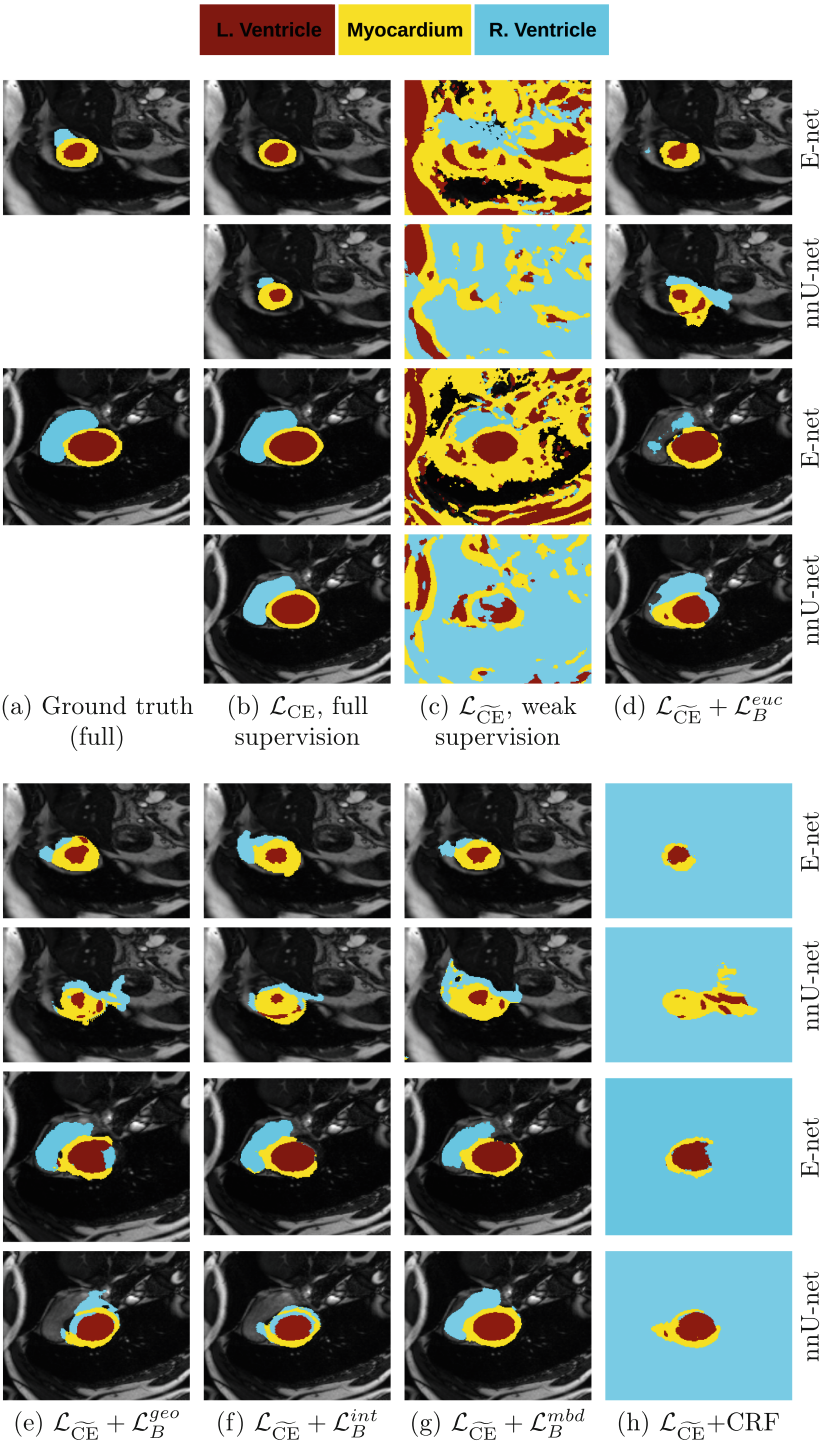


Fig. 4. Two example ACDC test set outputs, per method and architecture.

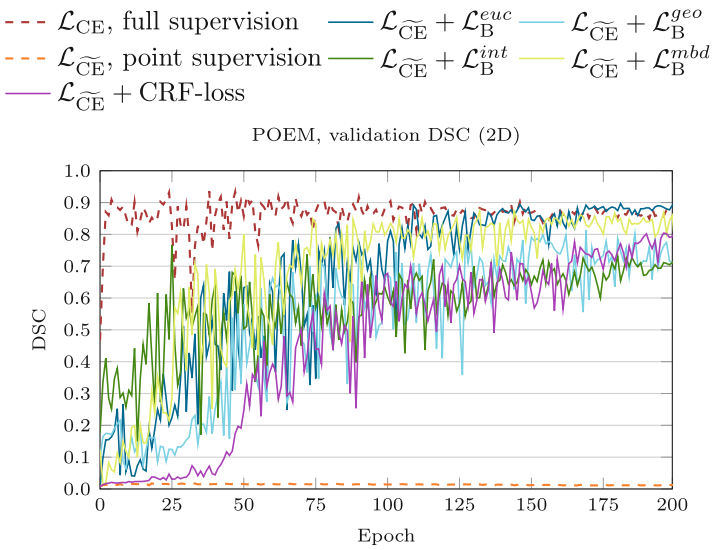


Fig. 5. Curve evolution of the average (foreground) validation 2D-Dice scores during training E-Net on POEM data. Losses $\mathcal{L}_{\dot{B}}$ use 2D-computed distance maps.

methods. Using Euclidean or MBD maps reach full-supervision scores, surpassing the other methods. However, due to the long computation times on 3D data from the POEM cohort, the curves show the evolution of the 2D Dice, which is less representative of the true success of the methods.

Using 3D Distance Maps. As expected, the results (Table 3) generally improve when training on 3D-computed distance maps, confirming that for 3D datasets with more complex class coocurrences distance maps should preferably be calculated in 3D directly (however, this can incur large preprocessing computational costs, see the supplementary material. Using E-Net, nost notable are decreases in HD95 values, as using volume-calculated distance maps provides more global information and additionally penalizes spatially unreasonable segmentations. The boundary loss based methods, particularly $\mathcal{L}_{\mathrm{B}}^{geo}$, are now able to compete with the CRF-loss. The validation curve evolution for training on 3D distance maps is shown in Fig. 6. Comparing it to the one with using 2D-computed distance maps (Fig. 5) we see that the curves for all the methods training with $\mathcal{L}_{\mathrm{B}}$ improve, with the exception of $\mathcal{L}_{\mathrm{B}}^{mbd}$ based one. The lack of improvement here could be attributed to the MBD bleeding through object boundaries (due to noise and/or lack of contrast) and propagating low distances further away in the volume, causing under-penalization. This is also suggested by the degradation in performance from 2D to 3D maps in Table 3. On the other hand, it allows for better segmentation of large and/or elongated (homogeneous) objects, which is also confirmed by improvement of liver segmentation scores in Table 3.

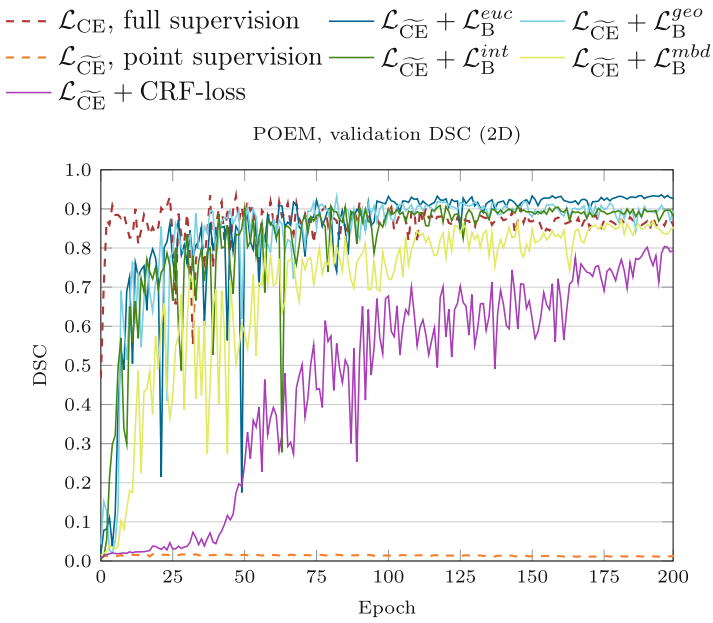


Fig. 6. Curve evolution of the average (over foreground classes) validation batch (2D) Dice scores during training, on the POEM dataset in multi-label segmentation training with E-Net. The boundary losses $\mathcal{L}_B$ use distance maps calculated on 3D volumes. The curves for training with $\mathcal{L}_{\text{CE}}$, $\mathcal{L}_{\widetilde{\text{CE}}}$ and CRF-loss are plotted again for easier comparison.

Table 3. Mean ↑DSC and ↓HD95 values over 3 independent runs with E-Net, on the POEM test set 3D volumes. Labels BLD, KDR, LVR, PNC, SPL and KDL stand for bladder, right kidney, liver, pancreas, spleen and left kidney respectively. For distribution boxplots over one run see the supplementary.

Method	BLD	KDR	LVR	PNC	SPL	KDL	ALL
$\mathcal{L}_{\text{CE}}$ (fully supervised)	↑ 0.607 ↓ 7.161	↑ 0.734 ↓ 4.261	↑ 0.895 ↓ 4.992	↑ 0.327 ↓ 10.504	↑ 0.588 ↓ 10.138	↑ 0.656 ↓ 3.630	↑ 0.687 ↓ 5.812
$\mathcal{L}_{\widetilde{\text{CE}}}$ (point annotations)	↑ 0.004 ↓ 106.589	↑ 0.015 ↓ 99.163	↑ 0.169 ↓ 71.601	↑ 0.005 ↓ 102.475	↑ 0.027 ↓ 112.903	↑ 0.024 ↓ 102.385	↑ 0.035 ↓ 85.017
w/$\mathcal{L}_{\text{B}}^{euc}$ (in 2D)	↑ 0.482 ↓ 9.576	↑ 0.689 ↓ 6.952	↑ 0.087 ↓ 25.049	↑ 0.436 ↓ 6.832	↑ 0.530 ↓ 7.673	↑ 0.664 ↓ 6.538	↑ 0.555 ↓ 8.946
w/$\mathcal{L}_{\text{B}}^{euc}$ (in3D)	↑ 0.471 ↓ 5.469	↑ 0.722 ↓ 4.028	↑ 0.363 ↓ 13.576	↑ 0.412 ↓ 5.417	↑ 0.622 ↓ 4.864	↑ 0.663 ↓ 3.094	↑ 0.608 ↓ 5.207
w/$\mathcal{L}_{\text{B}}^{geo}$ (in 2D)	↑ 0.354 ↓ 21.926	↑ 0.558 ↓ 12.403	↑ 0.078 ↓ 33.166	↑ 0.326 ↓ 13.904	↑ 0.323 ↓ 26.795	↑ 0.492 ↓ 8.188	↑ 0.447 ↓ 16.626
w/$\mathcal{L}_{\text{B}}^{geo}$ (in 3D)	↑ 0.475 ↓ 6.857	↑ 0.715 ↓ 4.948	↑ 0.391 ↓ 14.342	↑ 0.415 ↓ 6.212	↑ 0.673 ↓ 8.500	↑ 0.684 ↓ 2.878	↑ 0.622 ↓ 6.248
w/$\mathcal{L}_{\text{B}}^{int}$ (in 2D)	↑ 0.256 ↓ 57.424	↑ 0.571 ↓ 15.279	↑ 0.052 ↓ 25.584	↑ 0.322 ↓ 19.557	↑ 0.330 ↓ 68.651	↑ 0.409 ↓ 23.349	↑ 0.420 ↓ 29.978
w/$\mathcal{L}_{\text{B}}^{int}$ (in 3D)	↑ 0.483 ↓ 5.596	↑ 0.670 ↓ 8.650	↑ 0.611 ↓ 16.394	↑ 0.408 ↓ 6.234	↑ 0.618 ↓ 13.020	↑ 0.630 ↓ 5.846	↑ 0.631 ↓ 7.963
w/$\mathcal{L}_{\text{B}}^{mbd}$ (in 2D)	↑ 0.468 ↓ 7.716	↑ 0.634 ↓ 16.368	↑ 0.218 ↓ 29.664	↑ 0.386 ↓ 11.425	↑ 0.530 ↓ 20.234	↑ 0.598 ↓ 7.755	↑ 0.547 ↓ 13.309
w/$\mathcal{L}_{\text{B}}^{mbd}$ (in 3D)	↑ 0.466 ↓ 7.696	↑ 0.647 ↓ 4.415	↑ 0.360 ↓ 22.900	↑ 0.332 ↓ 8.855	↑ 0.496 ↓ 12.700	↑ 0.574 ↓ 4.648	↑ 0.553 ↓ 8.745
w/CRF-loss [29]	↑ 0.396 ↓ 8.835	↑ 0.685 ↓ 6.413	↑ 0.758 ↓ 19.622	↑ 0.448 ↓ 5.738	↑ 0.695 ↓ 8.316	↑ 0.661 ↓ 5.657	↑ 0.663 ↓ 7.797

Both Table 1 and 3 show that the proposed use of boundary loss with intensity-aware distances generally outperforms its original formulation under point supervision. However, the HD95 metric seemingly favours the $\mathcal{L}_{\text{B}}^{euc}$ setting. This is due to a smoother and more spatially contained output compared to using purely intensity-based distances that can result in more fragmented segmentations.

Training with CRF-loss may perform well, but specially designed background labels would potentially be needed to steer the CRF-loss training in the right direction. In addition, it incurs longer training times, see Supplementary C.

6 Conclusion and Future Work

We presented a novel approach of using intensity-aware distance with boundary loss to train CNNs for segmentation tasks under very weak supervision. Despite its simplicity, we achieve reasonable results without additional tuning or increased training time. Across multi-class segmentation tasks, our approach performs better or on-par compared to the state of the art CRF-loss that typically requires heavy tuning and is highly sensitive to parameter settings. In addition, it is more easily understood, as it is based on visually interpretable distance maps that have certain expected behaviours depending on the type of data. Being directly interpretable and easily applied across datasets, it provides a promising alternative to the CRF-loss training and methods derived from it.

Many small adaptations can be explored for further improvements. For example, intensity averaging over the annotation (or its border) prior to distance computation or adaptively controlling the spatial vs. intensity component contributions in distance definitions during training. In addition, combining the intensity-aware boundary loss training with CRF postprocessing remains to be investigated. Moreover, while we focused on intensity-aware distances that account for the underlying intensities in a direct way, texture-type distances could potentially further stabilize the training and prevent bleedout.

Acknowledgements. EB was partially funded by the Centre for interdisciplinary mathematics (CIM), Uppsala University. HK and MdB were funded by the Dutch Research Council (NWO), VI.C.182.042.

References

1. Asad, M., Dorent, R., Vercauteren, T.: FastGeodis: fast generalised geodesic distance transform. arXiv preprint arXiv:2208.00001 (2022)
2. Bai, X., Sapiro, G.: Geodesic matting: a framework for fast interactive image and video segmentation and matting. Int. J. Comput. Vis. **82**, 113–132 (2009). https://doi.org/10.1007/s11263-008-0191-z
3. Bearman, A., Russakovsky, O., Ferrari, V., Fei-Fei, L.: What's the point: semantic segmentation with point supervision. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9911, pp. 549–565. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46478-7_34
4. Bernard, O., et al.: Deep learning techniques for automatic MRI cardiac multi-structures segmentation and diagnosis: is the problem solved? IEEE Trans. Med. Imaging **37**(11), 2514–2525 (2018)
5. Chen, Z., et al.: Weakly supervised histopathology image segmentation with sparse point annotations. IEEE J. Biomed. Health Inform. **25**(5), 1673–1685 (2020)
6. Criminisi, A., Sharp, T., Blake, A.: GeoS: geodesic image segmentation. In: ECCV 2008, pp. 99–112 (2008)

7. Dai, J., He, K., Sun, J.: BoxSup: exploiting bounding boxes to supervise convolutional networks for semantic segmentation. In: 2015 IEEE International Conference on Computer Vision (ICCV), pp. 1635–1643 (2015). https://doi.org/10.1109/ICCV.2015.191
8. Dubost, F., et al.: Weakly supervised object detection with 2D and 3D regression neural networks. Med. Image Anal. **65**, 101767 (2020). https://doi.org/10.1016/j.media.2020.101767
9. Fan, J., Zhang, Z., Song, C., Tan, T.: Learning integral objects with intra-class discriminator for weakly-supervised semantic segmentation. In: Conference on Computer Vision and Pattern Recognition (CVPR) (2020)
10. Gulshan, V., Rother, C., Criminisi, A., Blake, A., Zisserman, A.: Geodesic star convexity for interactive image segmentation. In: 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 3129–3136 (2010). https://doi.org/10.1109/CVPR.2010.5540073
11. Isensee, F., Jaeger, P.F., Kohl, S.A., Petersen, J., Maier-Hein, K.H.: nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. Nat. Methods **18**(2), 203–211 (2021)
12. Ji, Z., Shen, Y., Ma, C., Gao, M.: Scribble-based hierarchical weakly supervised learning for brain tumor segmentation. In: Medical Image Computing and Computer Assisted Intervention – MICCAI 2019, pp. 175–183 (2019)
13. Kervadec, H., Bouchtiba, J., Desrosiers, C., Granger, E., Dolz, J., Ayed, I.B.: Boundary loss for highly unbalanced segmentation. Med. Image Anal. **67**, 101851 (2021)
14. Kervadec, H., Dolz, J., Tang, M., Granger, E., Boykov, Y., Ben Ayed, I.: Constrained-CNN losses for weakly supervised segmentation. Med. Image Anal. **54**, 88–99 (2019)
15. Kervadec, H., Dolz, J., Wang, S., Granger, E., Ayed, I.B.: Bounding boxes for weakly supervised segmentation: global constraints get close to full supervision. In: Medical Imaging with Deep Learning, pp. 365–381. PMLR (2020)
16. Kim, B., Jeong, J., Han, D., Hwang, S.J.: The devil is in the points: weakly semi-supervised instance segmentation via point-guided mask representation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 11360–11370 (2023)
17. Krähenbühl, P., Koltun, V.: Efficient inference in fully connected CRFs with Gaussian edge potentials. In: Advances in Neural Information Processing Systems 24, pp. 109–117. Curran Associates, Inc. (2011)
18. Lin, D., Dai, J., Jia, J., He, K., Sun, J.: ScribbleSup: scribble-supervised convolutional networks for semantic segmentation. In: Computer Vision and Pattern Recognition (CVPR), pp. 3159–3167 (2016)
19. Lind, L.: Relationships between three different tests to evaluate endothelium-dependent vasodilation and cardiovascular risk in a middle-aged sample. J. Hypertens. **31**, 1570–1574 (2013). https://doi.org/10.1097/HJH.0b013e3283619d50
20. Liu, W., He, Q., He, X.: Weakly supervised nuclei segmentation via instance learning. In: 2022 IEEE 19th International Symposium on Biomedical Imaging (ISBI), pp. 1–5. IEEE (2022)
21. Ma, J., et al.: How distance transform maps boost segmentation CNNs: an empirical study. In: Medical Imaging with Deep Learning. Proceedings of Machine Learning Research, vol. 121, pp. 479–492. PMLR (2020). https://proceedings.mlr.press/v121/ma20b.html

22. Mortazi, A., Khosravan, N., Torigian, D.A., Kurugol, S., Bagci, U.: Weakly supervised segmentation by a deep geodesic prior. In: Suk, H.I., Liu, M., Yan, P., Lian, C. (eds.) Machine Learning in Medical Imaging, pp. 238–246. Springer, Cham (2019)
23. Ngoc, M.Õ.V., Boutry, N., Fabrizio, J., Géraud, T.: A minimum barrier distance for multivariate images with applications. Comput. Vis. Image Underst. **197–198**, 102993 (2020). https://doi.org/10.1016/j.cviu.2020.102993
24. Paszke, A., Chaurasia, A., Kim, S., Culurciello, E.: ENet: a deep neural network architecture for real-time semantic segmentation (2016). https://doi.org/10.48550/ARXIV.1606.02147
25. Qu, H., et al.: Weakly supervised deep nuclei segmentation using points annotation in histopathology images. In: Medical Imaging with Deep Learning. Proceedings of Machine Learning Research, vol. 102, pp. 390–400. PMLR (2019). https://proceedings.mlr.press/v102/qu19a.html
26. Rajchl, M., et al.: DeepCut: object segmentation from bounding box annotations using convolutional neural networks. IEEE Trans. Med. Imaging **36**(2), 674–683 (2017). https://doi.org/10.1109/TMI.2016.2621185
27. Strand, R., Ciesielski, K.C., Malmberg, F., Saha, P.K.: The minimum barrier distance. Comput. Vis. Image Underst. **117**(4), 429–437 (2013). Special Issue on Discrete Geometry for Computer Imagery
28. Tang, M., Djelouah, A., Perazzi, F., Boykov, Y., Schroers, C.: Normalized cut loss for weakly-supervised CNN segmentation. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 1818–1827 (2018). https://doi.org/10.1109/CVPR.2018.00195
29. Tang, M., Perazzi, F., Djelouah, A., Ben Ayed, I., Schroers, C., Boykov, Y.: On regularized losses for weakly-supervised CNN segmentation. In: European Conference on Computer Vision (ECCV), Part XVI, pp. 524–540 (2018)
30. Toivanen, P.J.: New geodesic distance transforms for gray-scale images. Pattern Recogn. Lett. **17**(5), 437–450 (1996). https://doi.org/10.1016/0167-8655(96)00010-4
31. Wang, G., et al.: DeepIGeoS: a deep interactive geodesic framework for medical image segmentation. IEEE Trans. Pattern Anal. Mach. Intell. **41**(7), 1559–1572 (2019). https://doi.org/10.1109/TPAMI.2018.2840695
32. Xu, J., Schwing, A.G., Urtasun, R.: Learning to segment under various forms of weak supervision. In: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3781–3790 (2015). https://doi.org/10.1109/CVPR.2015.7299002
33. Yao, J., et al.: Position-based anchor optimization for point supervised dense nuclei detection. Neural Netw. **171**, 159–170 (2024)
34. Zheng, S., et al.: Conditional random fields as recurrent neural networks, pp. 1529–1537 (2015)
35. Zhou, Y., et al.: Prior-aware neural network for partially-supervised multi-organ segmentation. In: 2019 IEEE/CVF International Conference on Computer Vision (ICCV), pp. 10671–10680 (2019). https://doi.org/10.1109/ICCV.2019.01077

SpecSlice-ConvLSTM:Medical Hyperspectral Image Segmentation Using Spectral Slicing and ConvLSTM

Ming Hu[1,2,3,4], Jianfu Yin[1,3,4], Jing Wang[2(✉)], Yuqi Wang[1,3,4], Bingliang Hu[1,3,4(✉)], and Quan Wang[1,3,4(✉)]

[1] Key Laboratory of Spectral Imaging Technology, Xi'an Institute of Optics and Precision Mechanics (XIOPM), Chinese Academy of Sciences, Xi'an, China
{hbl,wangquan}@opt.ac.cn
[2] School of Printing, Packaging and Digital Media, Xi'an University of Technology, Xi'an 710048, China
wangjing63@xaut.edu.cn
[3] University of Chinese Academy of Sciences, Beijing, China
[4] Key Laboratory of Biomedical Spectroscopy of Xi'an, Xi'an Institute of Optics and Precision Mechanics (XIOPM), Chinese Academy of Sciences, Xi'an, China

Abstract. Medical hyperspectral imaging (MHSI) is a highly promising technology, offering new opportunities for computational pathology and precision medicine. However, the high spatial-spectral dimensions make simultaneous consideration of spatial and spectral features for image segmentation extremely challenging. In this study, we propose a segmentation network that slices the high-dimensional features of hyperspectral images into low-dimensional sequences, thus analyzing the spectral characteristics of medical hyperspectral images from a sequence perspective. This network is capable of learning both spectral and spatial information simultaneously, thereby enhancing the performance of image segmentation tasks. Its uniqueness lies in leveraging the Convolutional Long Short-Term Memory (ConvLSTM) structure to convert the long-range spectral dependencies of MHSI into relationships between low-channel image sequences, significantly improving the model's inference speed. Experiments conducted on the publicly available Multi-Dimensional Choledoch dataset demonstrate that compared to the state-of-the-art medical hyperspectral image segmentation algorithm, the Dual-Stream algorithm, our approach improved the DSC metric by 0.24%, increased inference speed by 1.4 times, and was 5–20 times faster than existing 3D networks.

Keywords: Medical hyperspectral images · MHSI segmentation

1 Introduction

Medical hyperspectral imaging (MHSI) is an advanced imaging technique that combines the principles of spectroscopy and imaging to acquire high-resolution

A. Antonacopoulos et al. (Eds.): ICPR 2024, LNCS 15313, pp. 211–225, 2025.
https://doi.org/10.1007/978-3-031-78201-5_14

spectral information from the surface of objects. This technology enables simultaneous acquisition of spatial and spectral information from objects, providing a new perspective and means for medical diagnosis, biomedical research, and medical imaging. Typically, MHSI is represented in the form of a hypercube, which includes hundreds of narrowband contiguous spectral bands and thousands of pixels.

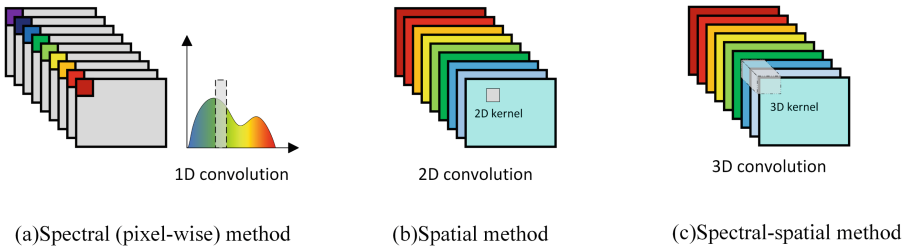


Fig. 1. Three convolutional neural network-based methods for hyperspectral image analysis. (a) Spectral (pixel-level) method analyzes individual spectral features independently through one-dimensional convolution. (b) Spatial method utilizes two-dimensional convolution to extract spatial features for data analysis. (c) Spectral-Spatial method performs three-dimensional convolution, fully utilizing the spatial and spectral information of the hypercube. Gray areas (a–c) are represented by 1D, 2D, and 3D kernels respectively, used for convolution operations. [13]

In recent years, Convolutional Neural Networks (CNNs), as one of the deep learning methods, have emerged as the most successful and popular image analysis approach [13–15] CNN models for medical hyperspectral images can analyze input data in spectral (1D), spatial (2D), and spectral-spatial (3D) methods (Fig. 1). Spectral models, also known as pixel-level models, analyze spectral profiles without considering spatial features. Since spectral information is a mixture of spectra profiles of various biological molecules, important features can be directly extracted from spectral information. To analyze the features of one-dimensional spectral signals, one-dimensional convolution can be applied (Fig. 1a).

While spectral methods offer promising prospects, they overlook spatial features. Malignant or abnormal tissues often exhibit irregular shapes and fuzzy edges, indicating that tissue morphology is also an important clinical feature. To obtain clinically meaningful outputs from tissue morphology, researchers employed spatial models to train CNN models using two-dimensional images at different wavelengths as input data (Fig. 1b).

The spectral-spatial model fully harnesses all the information from hyperspectral data (Fig. 1c). Three-dimensional convolution is typically employed in three-dimensional imaging techniques like magnetic resonance imaging [2] and computed tomography [7]. For hyperspectral data, three-dimensional convolution involves scanning convolutional kernels across the three-dimensional space,

such as along the x, y, and spectral axes. This method utilizes the entirety of hyperspectral data to train CNN models, resulting in output encompassing all tissue morphologies and biochemical features. However, compared to spectral and spatial models, three-dimensional convolution demands more computational resources and a richer set of hyperspectral data.

In order to simultaneously learn the spatial and spectral features of MHSIs, we transformed MHSIs into low-dimensional data sequences composed of every three adjacent spectral bands. We utilized these low-dimensional data sequences to focus on learning spatial features and the spectral features of the sequences. Additionally, to learn the global spectral features, we converted the intrinsic spectral features of MHSIs into long-range dependency relationships among sequences and employed a ConvLSTM structure to model these relationships between sequences. Our architecture is based on a U-shaped 2D CNN design, incorporating both the ResNet-34 [5] network structure and the ConvLSTM structure within the U-shaped framework. Compared to the 3D network structure of SpecTr [15], we employ a 2D U-Net [9] structure as the base framework, thereby reducing computational resource consumption. Compared to the Dual-Stream algorithm proposed by B Yun [14], we utilize a 2D network structure to jointly learn the spatial and spectral features of MHSIs, considering the correlation between these features.

2 Related Work

Hyperspectral imaging integrates imaging technology with spectroscopy, covering a continuous spectral range and enabling the scanning of multiple spectral bands. Unlike traditional RGB and grayscale images, hyperspectral imaging provides richer spectral bands and higher resolution, facilitating the detection of subtle spectral variations in invisible objects across diverse pathological conditions. To harness the spectral information embedded in three-dimensional hyperspectral data, Wang et al. [12] introduced Hyper-Net, a 3D fully convolutional network designed for segmenting melanoma in hyperspectral pathological images. Their approach includes a dual-pathway strategy in the encoding phase and employs dilated convolutions to capture fine features that might be lost in deeper layers, resulting in significantly enhanced segmentation accuracy.

In subsequent research, by embedding a transformer in the encoding part of U-Net [15] and applying it to image segmentation, dense correlations between bands can be learned. It inherits the advantages of Transformer and U-Net, making it more capable of segmenting medical images. However, the obtained information is easily influenced by irrelevant bands. Therefore, a sparse scheme was introduced to form the spectral transformer SpecTr, and experimental results showed that this scheme outperformed 3D U-Net and 2D U-Net.

In the latest research, B Yun et al. [14] proposed an accurate and fast medical hyperspectral image segmentation method based on factorized space and spectrum. This method utilizes the low-rank prior of MHSIs, exhibiting computational efficiency and plug-and-play capabilities, and can easily be inserted into any 2D architecture, greatly speeding up the network's inference speed.

3 Method

In mathematical terms, let $Z \in \mathbb{R}^{C \times H \times W}$ represent a three-dimensional volume of a pathological MHSI , where $H \times W$ is the spatial resolution and C is the number of channels in each hyperspectral image in the hyperspectral image dataset (i.e., the original spectral channel count). The objective of MHSI segmentation is to predict the predicted values of annotation labels for each pixel $\hat{Y} \in \{0,1\}^{H \times W}$. Our training set is denoted as $D = \{(Z_i, Y_i)\}_{i=1}^{N}$, where Y_i represents the ground truth values of each pixel of MHSI Z_i.

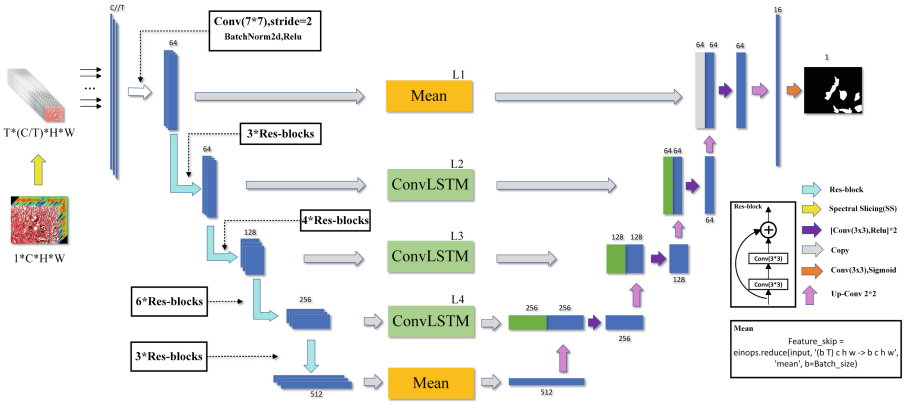


Fig. 2. The proposed SS-ConvLSTM architecture involves temporal decomposition of MHSI, followed by input to a 2D U-shaped network

The overall framework of our proposed method is illustrated in Fig. 2 where Fig. 3(a–b) details the process of Spectral Slicing, this process groups images of adjacent three bands from hyperspectral data, forming multiple three-channel 2D images, and applies visualization processing. This approach effectively reduces the complexity of the input MHSIs, facilitating subsequent feature extraction and segmentation. When optimizing the network architecture, we enhanced the U-Net framework by incorporating ResNet-34 as the core for feature extraction. ResNet-34 not only deepens the network but also addresses the gradient vanishing problem through residual connections, thereby improving the accuracy and efficiency of feature extraction.

What's more unique is that we introduced the ConvLSTM structure at the skip connections of the U-Net structure. Compared to direct connections in the original U-Net, the ConvLSTM structure can nonlinearly process image features extracted through Res-blocks. By stacking multiple ConvLSTMCell structures, we can capture long-distance dependency relationships between sequences in the image sequence sliced by spectral bands, which represents the global spectral information of MHSIs.

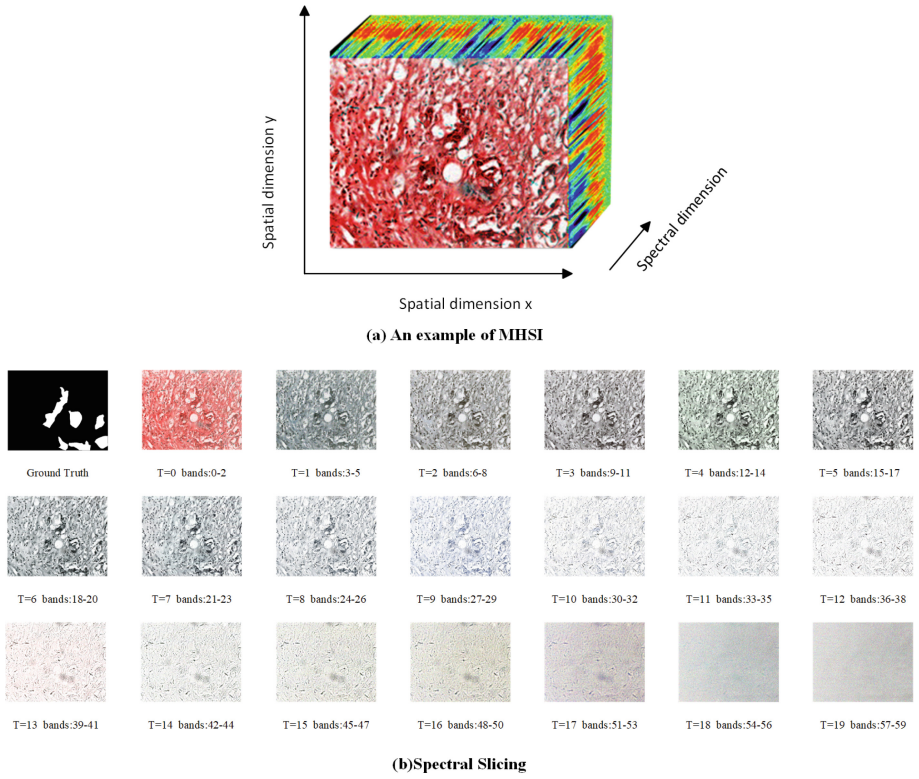


Fig. 3. (a) An example of MHSI (b) Visual pseudo-color image after spectral slicing of hyperspectral image

3.1 Spectral Slicing (SS)

We first perform Spectral Slicing (SS) on the hyperspectral image, reshaping it from the original form $Z \in \mathbb{R}^{1\times \mathrm{C}\times H\times W}$ to $Z \in \mathbb{R}^{T\times(\frac{C}{T})\times H\times W}$ (see Fig. 3(a–b)). In this process, we combine adjacent $\frac{C}{T}$ spectral bands to form T low-dimensional image combinations $x_0, x_1, ..., x_{T-1}$, treating each combination as a sample. Inspired by the spatial redundancy between adjacent spectral bands in medical hyperspectral imagery and the methods for temporal processing of natural image video sequences, the samples $x_0, x_1, ..., x_{T-1}$ in the sequence exhibit specific long-distance dependency relationships, corresponding to the intrinsic spectral features of the original medical hyperspectral image. We successfully transform the complex spectral features of the hyperspectral image into sequential relationships between low-dimensional image sequences. To fully capture the information between these sequences, we introduce the ConvLSTM structure. ConvLSTM effectively handles image data with spatial structures in addition to traditional sequential data. Through its internal convolution operations, it can preserve spatial information while capturing temporal dependencies.

3.2 ConvLSTM Structure

The ConvLSTM neural network cell (Fig. 4) typically consists of input layers, hidden layers, and output layers. The input layer receives external input data, while the hidden layer processes and stores information through memory cells, forget gates (f_t), and output gates (o_t). ConvLSTM models have the advantage of avoiding gradient vanishing/exploding, making them powerful in handling long sequence data. The output layer generates the final prediction results. In the diagram, each arrow represents a computational step, and nodes represent neurons [10]. Compared to traditional neural networks, the ConvLSTM neural network model has unique structure and advantages. (Fig. 4) Firstly, ConvLSTM introduces gate mechanisms, allowing for selective information transmission, thereby addressing the long-term dependency problem. Secondly, ConvLSTM models have memory capabilities, enabling them to capture historical information for more accurate predictions. Additionally, ConvLSTM models also have the advantage of avoiding gradient vanishing/exploding, making them powerful in handling long sequence data.

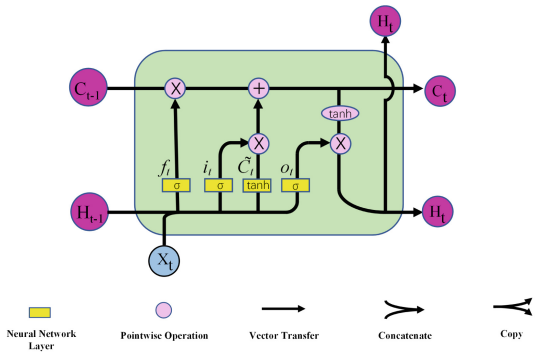


Fig. 4. The architecture of the ConvLSTMCell

The ConvLSTM can be expressed as follows:

$$i_t = \sigma(W_{xi} * X_t + W_{hi} * H_{t-1} + b_i) \tag{1}$$

$$f_t = \sigma(W_{xf} * X_t + W_{hf} * H_{t-1} + b_f) \tag{2}$$

$$o_t = \sigma(W_{xo} * X_t + W_{ho} * H_{t-1} + b_o) \tag{3}$$

$$\tilde{C}_t = \tanh(W_{xc} * X_t + W_{hc} * H_{t-1} + b_c) \tag{4}$$

$$C_t = f_t \odot C_{t-1} + i_t \odot \tilde{C}_t \tag{5}$$

$$H_t = o_t \odot \tanh(C_t) \tag{6}$$

where $*$ and $\odot$ denote convolution and Hadamard product respectively. X_t is the input tensor, H_t is the hidden state tensor, C_t is the memory cell tensor, W_{x*} and W_{h*} are the two-dimensional convolutional kernels corresponding to the input and hidden state, b_i, b_f, b_o, and b_c are the bias terms [1].

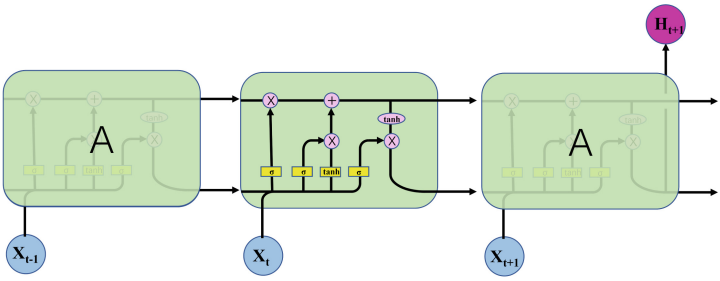


Fig. 5. The architecture of the classical ConvLSTM structure

The classical ConvLSTM structure is designed to take in multiple sequences as input and output the final hidden layer results (Fig. 5). However, for complex data such as MHSIs, the relationships between the low-dimensional data sequences formed after spectral slicing may be more subtle and intricate. Traditional ConvLSTM structures may not fully capture these relationships when dealing with such data, hence they have limitations.

Given the excellent performance of ConvLSTM structures in handling time series data, in order to more comprehensively capture the subtle relationships within hyperspectral images, we have designed two ConvLSTM structures that are better suited for medical hyperspectral images: Single Input Single Output Model (SISOM) and Dual Input Dual Output Model (DIDOM).

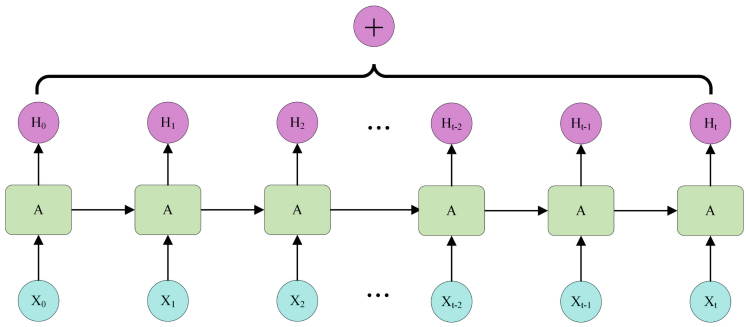


Fig. 6. The architecture of the Single Input Single Output Model (SISOM)

Single Input Single Output Model (SISOM). By observing the structure of the classical ConvLSTM, we found that it is limited due to the forgetfulness of the ConvLSTM structure, making it ineffective in capturing long-range dependencies. To fully utilize all the input information, we propose the forward Single Input Single Output Model (Fig. 6) the hidden states (H_t) from different time points are fused together through summation operation to fully utilize the information from different time steps, as shown in Eq. (7):

$$Feature_{skip} = \sum_{i=0}^{T-1} H_i \tag{7}$$

Dual Input Dual Output Model (DIDOM). The SISOM structure can only learn the sequence relationships in one direction, whereas theoretically, sequence relationships should be bidirectional. Therefore, to better capture the inherent dependencies between sequences, we designed a bidirectional input bidirectional output model. The uniqueness of this model lies in simultaneously feeding the original sequence forward and backward into a shared ConvLSTM structure. By leveraging the sequence processing capabilities of the ConvLSTM structure, we can more effectively learn the underlying relationships within the sequences obtained after spectral slicing.

It is worth noting that although this structure inputs forward and backward sequences into the ConvLSTM structure, its parameter count is comparable to that of the SISOM structure. However, to delve deeper into learning the relationships between sequences, this model requires more computational resources. This design aims to enhance the model's expressive power and predictive performance, enabling it to more comprehensively consider the features and patterns of sequence data, as shown Eq. (8):

$$Feature_{skip} = \sum_{i=0}^{T-1} H_i + \sum_{i=0}^{T-1} \overline{H}_i \tag{8}$$

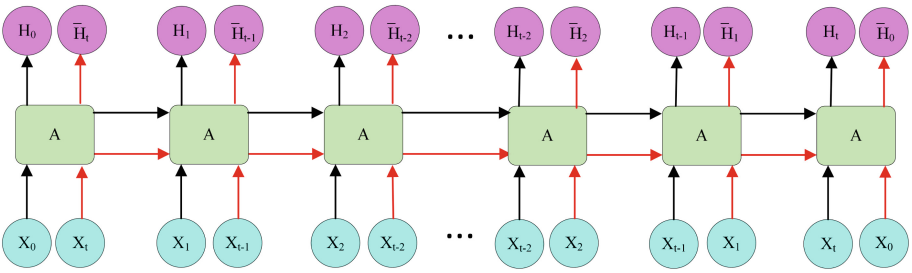


Fig. 7. The architecture of the Dual Input Dual Output Model (DIDOM)

As both our designed SISOM and DIDOM models are based on the ConvLSTM structure, our approach will be collectively referred to as SS-ConvLSTM in the following text. Only when it's necessary to distinguish and compare them, we will refer to them as SS-SISOM and SS-DIDOM.

4 Experimental Results

4.1 Dataset

MDC Dataset. Multi-Dimensional Choledoch (MDC) Dataset [14] comprising 538 scenes with high-quality labels for binary MHSI segmentation tasks. These MHSIs were collected using a hyperspectral system with a 20× objective lens, covering wavelengths from 550 nm to 1000 nm for MDC, resulting in 60 spectral bands per scene. The size of individual band images in the MDC dataset was resized to 256×320. The MDC dataset was partitioned into training, validation, and test sets using a patient-centric hard split approach with a ratio of 3:1:1.

4.2 Experimental Setup

We trained using an Adam optimizer with a combination of dice loss and cross-entropy loss for a batch size of 4 and 100 epochs. Segmentation performance was evaluated using Dice-Sørensen coefficient (DSC), Intersection of Union (IoU), and Hausdorff Distance (HD), Throughput (images per second), MACs (Multiply Accumulate Operations) and Params metrics. We utilized the PyTorch framework and three NVIDIA GeForce RTX 3090 GPUs for implementation.

4.3 Evaluation of the Proposed Strategy

Comparison Between Different Feature Extraction Backbones After Using Spectral Slicing. In Table 1, we used four structures from the ResNet series: ResNet-18, ResNet-34, ResNet-50, and ResNet-101 to explore the impact of different depths of ResNet architectures on the results and select the optimal feature extraction backbone. The results show that as the depth of ResNet increases, the model's performance improves on certain metrics. For example, ResNet-101 has the highest IOU and DSC, with values of 60.57 and 73.95, respectively, indicating better segmentation accuracy. On the other hand, as the network depth increases, the computational complexity and number of parameters also increase significantly. ResNet-101's MACs and parameter count are 263.39G and 53.09M, respectively, which are much higher than other variants. Additionally, ResNet-101 has the slowest processing speed, with a throughput of only 17.27 images/s. Considering accuracy, computational complexity, and processing speed comprehensively, ResNet-34 shows balanced performance across various metrics and might be an ideal choice. It ensures high segmentation accuracy while maintaining relatively low computational complexity and parameter count, and it also has a higher inference speed. This indicates that ResNet-34 can provide the best balance between performance and efficiency when used as the U-Net backbone.

Table 1. Results of ResNet architectures with different depths.Performance comparison in "mean" in MDC dataset.

Backbone	**Method**	**SS**	**IOU** ↑	**DSC** ↑	**HD** ↓	**Throuput** ↑	**MACs(G)** ↓	**Params(M)** ↓
ResNet-18	U-Net	✓	59.05	72.38	81.47	42.72	63.39	15.90
ResNet-34	U-Net	✓	59.80	73.39	76.36	34.72	123.96	26.01
ResNet-50	U-Net	✓	60.47	73.76	80.30	25.82	141.5	34.1
ResNet-101	U-Net	✓	60.57	73.95	78.71	17.27	263.39	53.09

Table 2. The following is a comparison of the effects between our two improved ConvLSTM models and the classical ConvLSTM structure.Performance comparison in "mean(std)" in MDC dataset.

Backbone	Method	SS	Model	IOU ↑	DSC ↑	HD ↓
ResNet-34	U-Net	✓	Classical ConvLSTM	61.48 (18.23)	74.48 (14.85)	77.54 (32.04)
ResNet-34	U-Net	✓	SISOM	62.42 (16.70)	75.48 (13.56)	74.41 (29.96)
ResNet-34	U-Net	✓	DIDOM	62.71 (16.83)	75.68 (13.64)	76.59 (30.04)

Comparison of the Two Improved ConvLSTM Models Proposed in This Paper and the Classical ConvLSTM Structure. Table 2 illustrates the performance comparison in segmentation tasks between two improved ConvLSTM models proposed in this paper (SISOM and DIDOM) and the classical ConvLSTM structure. The results demonstrate that the enhanced SISOM and DIDOM models outperform the classical ConvLSTM in terms of IOU and DSC metrics, respectively showcasing higher segmentation accuracy. The SISOM structure shown in Fig. 6 and the DIDOM structure shown in Fig. 7, compared to the classic ConvLSTM structure in Fig. 5, can capture all hidden layer states. These hidden states capture the complex relationships between spectral slices, effectively characterizing abstract spectral features in hyperspectral data. The method proposed in this paper further enhances the model's representation capability of spectral information by integrating these hidden layer states through addition.

Ablation Study. Our SS-ConvLSTM model demonstrates a high degree of adaptability. Initially, we conducted a comprehensive ablation study to assess the effectiveness of each component. We use the Unet architecture and Resnet34 feature extraction blocks as the basic framework. By adopting the SS-ConvLSTM framework, we merged the SS module with the ConvLSTM structure. We inserted our designed ConvLSTM modules at different skip-connection points in the network architecture, labeled as L1, L2, L3, and L4, as illustrated in Fig. 2. When the ConvLSTM structure is not inserted at the skip connection points, we average the sequential features obtained in the encoding part and then concatenate them with the features obtained in the decoding part. The results of the ablation study in Table 3 indicate that by incorporating the spectral slicing

module, our model achieved a segmentation performance improvement of over 1.7% (73.39 vs. 71.68). The spectral slicing module divides MHSIs into multiple low-channel data, which helps reduce data dimensionality, mitigate overfitting risks, and enhance the model's generalization ability, thereby improving segmentation performance. Additionally, leveraging the ConvLSTM structure enables effective capture of long-range dependencies in sequential data, facilitating learning of spectral features in medical hyperspectral images. This compensates for the disruption of spectral features in medical hyperspectral images caused by spectral slicing, further enhancing the model's segmentation capability. We used an example to showcase the details of using the SS module and embedding our proposed ConvLSTM structure at different positions (Fig. 8).

Table 3. Performance comparison in "mean (std)" on MDC dataset.

SS	SISOM				IOU↑	DSC↑	HD↓
	L1	L2	L3	L4			
					57.88 (17.58)	71.68 (14.87)	77.42 (31.98)
✓					59.80 (16.89)	73.39 (13.85)	76.36 (31.22)
✓	✓				61.29 (16.73)	74.60 (13.61)	77.23 (31.55)
✓		✓			60.93 (16.66)	74.32 (13.69)	78.55 (32.17)
✓			✓		60.17 (17.38)	73.58 (14.44)	77.55 (30.11)
✓				✓	61.82 (18.55)	74.66 (15.38)	76.79 (31.08)
✓	✓	✓	✓	✓	60.43 (16.75)	73.87 (14.16)	78.03 (30.08)
✓	✓	✓			60.81 (18.07)	73.91 (15.45)	80.41 (33.60)
✓	✓		✓		61.82 (16.31)	75.08 (13.25)	75.25 (30.50)
✓		✓	✓	✓	**62.42 (16.70)**	**75.48 (13.56)**	**74.41 (29.96)**

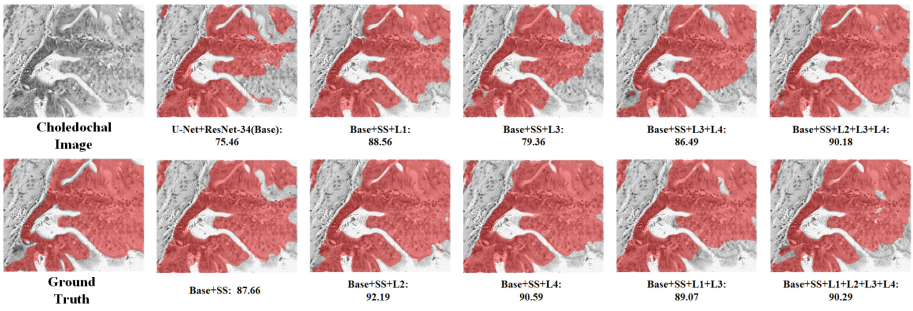


Fig. 8. Based on the U-Net+ResNet-34 architecture as the foundation, we use SS modules and insert our proposed SISOM structure at different skip connection positions (L1, L2, L3, L4) to demonstrate segmentation details. The evaluation metric in the figure is DSC.

Comparison with Other Medical Hyperspectral Image Segmentation Methods. In Table 4, we compared our performance with state-of-the-art (SOTA) methods, evaluating metrics including the "mean (std)" of IOU, DSC, and HD, as well as throughput, MACs, and Parameters (Params). The best results are highlighted. The experiments demonstrate that our proposed models (SS-SISOM and SS-DIDOM) outperform other algorithms in segmentation accuracy measured by DSC. Additionally, our models achieve throughputs of 23.50 images/s and 19.90 images/s, significantly higher than the state-of-the-art algorithms, Dual-Stream algorithm (13.84 images/s).

Table 4. Performance comparison with SOTA methods on the MDC dataset.

	Method	IOU(%)↑	DSC(%)↑	HD↓	Throughput↑	MACs(G)↓	Params(M)↓
2D	SS-DIDOM (ours)	**62.71 (16.83)**	**75.68 (13.64)**	76.59 (30.04)	19.90	305.24	32.21
	SS-SISOM(ours)	62.42 (16.70)	75.48 (13.56)	**74.41 (29.96)**	23.50	214.60	32.21
	Dual-Stream [14]	62.28 (16.25)	75.44 (13.31)	77.70 (30.34)	13.84	111.29	27.06
	DeepLabV3+ [3]	58.82 (18.40)	72.32 (15.27)	77.50 (33.94)	**43.10**	13.55	**22.62**
	FPN [8]	59.73 (18.48)	73.03 (15.37)	76.14 (32.36)	42.88	**12.26**	23.33
	U-Net [9]	57.88 (17.58)	71.68 (14.87)	77.42 (31.98)	39.93	13.48	24.62
3D	3D-UNet [4]	59.08 (18.02)	72.55 (15.37)	82.40 (28.91)	4.04	1110.77	-
	nnUNet [6]	60.49 (15.80)	74.12 (12.91)	79.87 (30.50)	1.92	1253.82	-
	HyperNet [12]	58.86 (17.75)	72.47 (14.77)	83.75 (33.85)	0.99	1512.24	-
	Swin-UNETR [11]	58.54 (16.30)	72.39 (14.31)	78.38 (31.74)	1.45	245.04	-
	SpecTr [15]	59.99 (16.12)	73.66 (13.30)	76.92 (31.93)	1.40	1049.72	-

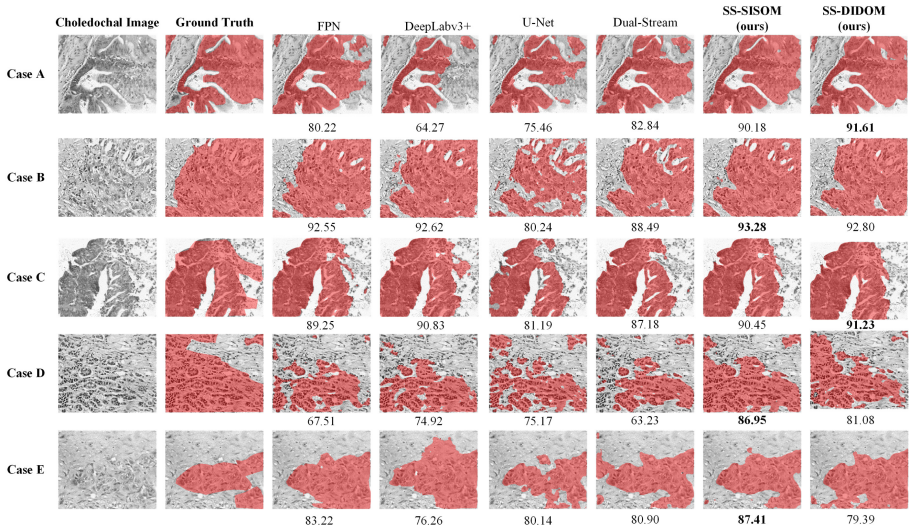


Fig. 9. Qualitative visualizations of our proposed methods and other methods on the MDC dataset. The evaluation metric in the figure is DSC.

Through qualitative visualizations on the MDC dataset (Fig. 9), we present the performance of the proposed SS-ConvLSTM method and other methods. The qualitative visualizations showcase the segmentation results of the SS-ConvLSTM method alongside other popular segmentation methods. Through these comparisons, we demonstrate the advantages of the SS-ConvLSTM method in capturing fine structures and boundary information in the images. These qualitative visualizations provide an intuitive understanding of the performance of different methods, aiding researchers and medical professionals in assessing the applicability and accuracy of various approaches in real-world applications. These results also serve to guide future research and development in the field of medical image analysis.

5 Discussion

Our method introduces the ConvLSTM module into the U-Net architecture to enhance model performance. Despite the additional modules increasing the model's parameters and computational load, resulting in a decrease in throughput metrics, we observed a significant improvement in the model's segmentation capability. Accurate segmentation is crucial in medical image processing. Therefore, we believe that sacrificing some performance for better segmentation is acceptable. Additionally, we believe that further optimization and adjustments can mitigate the decrease in throughput metrics to achieve a better balance while maintaining high segmentation accuracy. It is worth noting that the FPN and DeepLabV3+ structures we compare also utilize ResNet-34 as the backbone network for extracting features from hyperspectral images. Thus, they are comparable to the U-Net structure based on ResNet-34 in terms of parameter count, computational load, and throughput metrics.

Experiment with the Results of Different Numbers of Band Combinations. The method in the text only processes the 60 bands in the MDC dataset by slicing every 3 adjacent bands. However, for datasets with 200 or more bands, this method leads to significant redundancy. Therefore, in the SS-DIDOM method, we investigated the effect of combining different numbers of adjacent bands on the experimental outcomes. The Table 5 demonstrates the impact of combining different numbers of bands on various performance metrics in the SS-DIDOM method. When combining 3 bands, the IOU and DSC values are the highest, at 62.71 and 75.68 respectively, indicating the best segmentation performance. As the number of bands increases, the throughput improves, reaching a maximum of 34.75 (with 12 bands), while the computational complexity (MACs) significantly decreases, with a minimum of 82.06 (with 12 bands). Overall, combining more bands can enhance processing speed and reduce computational complexity, but the best segmentation performance is achieved with a combination of 3 bands. This suggests that when handling hyperspectral image data with more bands, combining a larger number of spectral bands can reduce redundancy after slicing and accelerate inference speed.

Table 5. The results of using the SS-DIDOM method with different numbers of adjacent band combinations.

SS-DIDOM					
bands	IOU ↑	DSC ↑	HD ↓	Throughput ↑	MACs(G) ↓
2	62.01(17.06)	75.08(14.17)	77.13(35.24)	14.62	454.03
3	**62.71 (16.83)**	**75.68 (13.64)**	76.59 (30.04)	19.90	305.24
4	60.88(16.13)	74.37(13.15)	76.01(29.66)	22.98	230.85
5	61.59(16.74)	74.82(13.72)	**75.87(30.65)**	25.72	186.21
6	60.81(16.95)	74.19(13.79)	76.72(32.28)	27.23	156.45
10	59.55(17.37)	73.08(14.59)	78.36(32.44)	33.25	96.94
12	60.81(19.13)	73.76(15.88)	82.37(32.42)	**34.75**	**82.06**

6 Conclusion

We decomposed medical hyperspectral images into multiple low-channel data through spectral slicing, and processed them using the ConvLSTM structure. This transformation converts the long-range spectral dependencies of hyperspectral images into relationships between low-channel image sequences. Testing on the MDC dataset showed that our approach improved the DSC metric by 0.24% and achieved an inference speed approximately 1.4 times faster. Compared to existing 3D algorithms, our method showed a speed enhancement of 5-20 times. Our approach provides a more efficient and accurate means of handling medical hyperspectral image segmentation. By jointly considering spectral and spatial features in the learning process, our method not only enhances segmentation accuracy but also significantly improves inference speed.

Acknowledgement. The research was supported by the Key Laboratory of Spectral Imaging Technology, Xi'an Institute of Optics and Precision Mechanics of the Chinese Academy of Sciences [grant number 54S18-014]; the Key Laboratory of Biomedical Spectroscopy of Xi'an [grant number 201805050ZD1CG34]; the Outstanding Award for Talent Project of the Chinese Academy of Sciences [grant number 29J20-052-III]; the National science basic research program of Shaanxi under Grant 2024JC-YBMS-552.

References

1. Azad, R., Asadi-Aghbolaghi, M., Fathy, M., Escalera, S.: Bi-directional convlstm u-net with densley connected convolutions. In: Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops (2019)
2. Chen, L., Wu, Y., DSouza, A.M., Abidin, A.Z., Wismüller, A., Xu, C.: MRI tumor segmentation with densely connected 3D CNN. In: Medical Imaging 2018: Image Processing, vol. 10574, pp. 357–364. SPIE (2018)

3. Chen, L.C., Zhu, Y., Papandreou, G., Schroff, F., Adam, H.: Encoder-decoder with atrous separable convolution for semantic image segmentation. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 801–818 (2018)
4. Çiçek, Ö., Abdulkadir, A., Lienkamp, S.S., Brox, T., Ronneberger, O.: 3D U-Net: learning dense volumetric segmentation from sparse annotation. In: Ourselin, S., Joskowicz, L., Sabuncu, M.R., Unal, G., Wells, W. (eds.) MICCAI 2016 Part II. LNCS, vol. 9901, pp. 424–432. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46723-8_49
5. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778 (2016)
6. Isensee, F., et al.: nnU-Net: self-adapting framework for u-net-based medical image segmentation. arXiv preprint arXiv:1809.10486 (2018)
7. Ker, J., Singh, S.P., Bai, Y., Rao, J., Lim, T., Wang, L.: Image thresholding improves 3-dimensional convolutional neural network diagnosis of different acute brain hemorrhages on computed tomography scans. Sensors **19**(9), 2167 (2019)
8. Kirillov, A., Girshick, R., He, K., Dollár, P.: Panoptic feature pyramid networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 6399–6408 (2019)
9. Ronneberger, O., Fischer, P., Brox, T.: U-Net: convolutional networks for biomedical image segmentation. In: Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F. (eds.) MICCAI 2015 Part III. LNCS, vol. 9351, pp. 234–241. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-24574-4_28
10. Shi, X., Chen, Z., Wang, H., Yeung, D.Y., Wong, W.K., Woo, W.C.: Convolutional LSTM network: a machine learning approach for precipitation nowcasting. In: Advances in Neural Information Processing Systems, vol. 28 (2015)
11. Tang, Y., et al.: Self-supervised pre-training of swin transformers for 3D medical image analysis. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 20730–20740 (2022)
12. Wang, Q., et al.: Identification of melanoma from hyperspectral pathology image using 3D convolutional networks. IEEE Trans. Med. Imaging **40**(1), 218–227 (2020)
13. Yoon, J.: Hyperspectral imaging for clinical applications. BioChip J. **16**(1), 1–12 (2022)
14. Yun, B., Li, Q., Mitrofanova, L., Zhou, C., Wang, Y.: Factor space and spectrum for medical hyperspectral image segmentation. In: Greenspan, H., et al. (eds.) MICCAI 2023. LNCS, vol. 14223, pp. 152–162. Springer, Cham (2023). https://doi.org/10.1007/978-3-031-43901-8_15
15. Yun, B., Wang, Y., Chen, J., Wang, H., Shen, W., Li, Q.: Spectr: spectral transformer for hyperspectral pathology image segmentation. arXiv preprint arXiv:2103.03604 (2021)

Advancing Brain Tumor Diagnosis: A Hybrid Approach Using Edge Detection and Deep Learning

Ha Anh Vu(✉) and Szilárd Vajda

Department of Computer Science, Central Washington University, Ellensburg, WA 98926, USA
{HaAnh.Vu,Szilard.Vajda}@cwu.edu
https://www.cwu.edu/academics/computer-science/

Abstract. Brain tumor classification from MRI scans demands precise image analysis, a challenge compounded by the variable morphology and location of tumors. Addressing this, our study presents an innovative approach that combines edge detection with a hierarchical deep learning framework to classify brain tumors accurately. This method enhances edge clarity, facilitating the deep learning model's ability to distinguish between meningioma, glioma, and pituitary tumors. By deploying a two-stage model, initially segregating a meta-tumor class and pituitary and subsequently refining the meta-tumor class into glioma and meningioma with a binary classifier, we capitalize on the strengths of both traditional image processing and advanced neural networks. The already proven ResNet50 architecture, our model's backbone, benefits from transfer learning, enabling efficient feature extraction from the edge image tailored to brain tumor recognition. Our results, evidenced by an over 96% overall accuracy rate obtained on a large benchmark brain tumor dataset, underscore the potential of integrating edge detection processing with deep learning. This integrative multi-level strategy promises to streamline the diagnostic process, offering a reliable, fast, and cost-effective solution that could reduce the need for expensive human specialist intervention.

Keywords: Brain Tumor · MRI · Canny Edge Detector · Deep Learning · Medical Image Analysis

1 Introduction

Medical image analysis and classification is an emerging technology in different medical domains, and it is used with huge success in fields such as cell counting [1], chest x-ray analysis [2,3], brain scan analysis [4,5], etc. All these approaches facilitate a non-invasive approach based on classical image processing techniques and recognition mechanisms to help health professionals make appropriate decisions without being concerned by the limited number of existing professionals,

A. Antonacopoulos et al. (Eds.): ICPR 2024, LNCS 15313, pp. 226–241, 2025.
https://doi.org/10.1007/978-3-031-78201-5_15

which are expensive to train and their possible fatigue during the analysis of such images which can lead to erroneous decisions and diagnoses.

Brain tumors encompass a variety of abnormal cellular growths within the brain (see Fig. 4, Fig. 5 and Fig. 6), each with potentially serious implications for patient health and treatment outcomes. Accurately identifying these tumors is paramount [6] in devising effective treatment strategies. Yet, the task is fraught with complexity due to the intricate anatomy of the brain and the subtle differences between tumor types. Traditionally, the diagnosis of brain tumors has relied heavily on the expertise of medical professionals analyzing magnetic resonance imaging (MRI) scans. While MRI provides detailed images of brain structures without exposing patients to ionizing radiation, the manual interpretation of these images is subjective. It can lead to inconsistencies, especially given the nuanced distinctions between various tumor appearances.

The advent of computer vision and machine learning technologies [4], especially deep learning [5], has introduced promising advancements in medical imaging analysis. These technologies offer the potential to automate and enhance the accuracy of brain tumor classification from MRI scans, addressing some of the inherent challenges of manual interpretation. However, despite these advances, applying such techniques to brain tumor detection is not without its hurdles [4,7]. Key issues include the requirement for large, annotated datasets for training these models, the selection of the appropriate information for training the models, the substantial computational resources needed, and the critical need for the resulting models to be both accurate and interpretable by medical professionals.

This paper proposes a novel approach to overcome these challenges, utilizing advanced image processing techniques coupled with a hierarchical dual-model classification system. By employing Canny edge detection for image preprocessing, we aim to improve the delineation of tumor boundaries, facilitating their subsequent identification. This choice is motivated by current strategies involving deep neural networks work mainly on the original MRI scans, which contain a multitude of unnecessary information. Meanwhile, our solution focuses on regions where tumors can be spotted. Implementing a two-tiered model architecture leverages deep learning capabilities to achieve more precise tumor classification in each decision-making step.

The rest of the paper is structured as follows: Sect. 2 gives an overview of the state-of-the-art, in Sect. 3. The proposed method will be described, –involving image processing, data augmentation, and classification by a deep neural network. Section 4 will present the computational setup, the benchmark data of the MRI brain scans, the metrics considered for the experiments, the results obtained, and last but not least, some comparisons with similar methods proposed in the literature. Finally, a brief summary highlighting the technique and the results will be provided in Sect. 5.

2 Related Work

Magnetic Resonance Imaging (MRI) and Deep Learning (DL) intersection for brain tumor detection and classification represent a rapidly evolving research

area [4]. Pioneering studies have leveraged the capabilities of convolutional neural networks (CNNs) [8], transfer learning, and image processing techniques to improve diagnostic accuracy [5,9]. However, critically examining these methodologies reveals persistent challenges that hinder their clinical applicability.

2.1 Deep Learning in MRI-Based Brain Tumor Classification

This section focuses into the evolving landscape of deep learning methodologies within MRI-based brain tumor classification.

Nazir et al. [10] offer a detailed review of DL in MRI-based brain tumor classification, focusing on the pivotal role of CNNs in driving the field forward. They identify crucial challenges, including the need for large, annotated datasets, high computational overhead, and models' interpretability issues, pointing towards efficient, transparent models capable of operating with limited data.

Expanding upon these observations, Sarfarazi et al. [11] research into the performance of specific CNN models such as AlexNet [12], VGG16 [13], and ResNet [14] in brain tumor identification. They emphasize the crucial impact of dataset quality on model performance and note the dependency on comprehensive, high-quality datasets for practical training and validation.

Khan et al. [15] investigate a dual-model strategy combining a 23-layer CNN with VGG16, achieving high classification accuracy. They highlight the importance of extensive datasets and the need for model interpretability, especially in clinical applications.

Innovatively addressing data scarcity, Soumik et al. [16] and Swati et al. [17] utilize transfer learning with InceptionV3 and VGG19, respectively, for tumor classification. By pre-training on ImageNet and employing strategic fine-tuning, they demonstrate how transfer learning can significantly enhance classification accuracy, even with limited datasets, setting a precedent for future diagnostic advancements.

Irmak [18] pushes the envelope further by applying CNNs for multi-classification of brain tumors, highlighting the importance of hyperparameter optimization through grid search to significantly boost accuracy, showcasing the critical role of fine-tuning in leveraging deep learning for medical diagnostics.

2.2 Addressing Data Quality and Computational Challenges

We examine methods aimed at refining tumor classification through enhanced data quality and computational efficiency. From image enhancement and clustering techniques to dense networks and support vector machine (SVM) approaches, the focus lies on preprocessing, segmentation accuracy, and scalable diagnostic tools.

Rasheed et al. [19] and Sahoo et al. [20] contribute to the discourse with methods that integrate image enhancement and clustering for refined tumor classification. They stress the significance of preprocessing and segmentation accuracy while acknowledging the need for image quality improvements and adaptable segmentation techniques.

Further contributions by Zhou et al. [21], Montoya et al. [22], and Dheepak et al. [23] explore dense networks and SVM approaches, offering solutions to computational efficiency and the necessity for large datasets through novel preprocessing strategies and multi-kernel SVM approaches for feature extraction, highlighting the ongoing evolution toward more precise and adaptable diagnostic tools.

Cheng et al. [24] innovate by augmenting tumor regions and partitioning them for a detailed analysis combining several feature extraction techniques, illustrating the potential of leveraging the tumor's spatial information to elevate classification accuracy.

2.3 Innovative Approaches and Frameworks

Innovative strides in overcoming limitations in conventional CNNs are explored in this section. From hybrid models like DeepTumorNet to novel techniques integrating spatial relationships and active contour algorithms, the emphasis is on enhancing tumor detection accuracy and reliability for automatic diagnosis.

The introduction of DeepTumorNet by Raza et al. [25] and an automated DL model by Ullah et al. [26] represents a leap in overcoming conventional CNN limitations through hybrid models and information fusion strategies, addressing the challenge of imbalanced datasets.

Afshar et al. [27] take a unique path by enhancing Capsule Networks with tumor boundary information, focusing on spatial relationships to improve network performance. In contrast, Shanaka et al. [28] combine deep learning with active contour algorithms for segmentation, presenting a method that significantly boosts tumor detection accuracy.

Building on previous research, it becomes clear that each study has strengths and limitations, such as limited data sets, complex and computationally intensive classification models, and modest performance figures. While Shanaka et al. [28] employed contouring methods to highlight tumor edges, their approach does not segregate tumors from other brain components as effectively as our method. Our work focuses on isolating the tumor from the edge image and classifying this specific region of the MRI scan. By emphasizing tumor edges, we effectively delineate brain tumors from other components, such as brain tissue and fluid, in 2D MRI images, significantly boosting model training efficiency and accuracy. Unlike the approach in [28], we rely on tumor contours for spatial representation and location. Additionally, our hierarchical classification model distinctively identifies three pathological tumor types: glioma, meningioma, and pituitary. Given the similarity between meningioma (see Fig. 5) and glioma (see Fig. 4), a multi-level classification scheme is necessary. In the first stage, we separate pituitary malformations from other pathologies. In the second stage, a dedicated classification scheme differentiates between glioma and meningioma tumors, enabling efficient and reliable automatic diagnosis.

3 Methodology

This section outlines our study's comprehensive approach to enhancing MRI-based brain tumor detection and classification. We describe the use of the

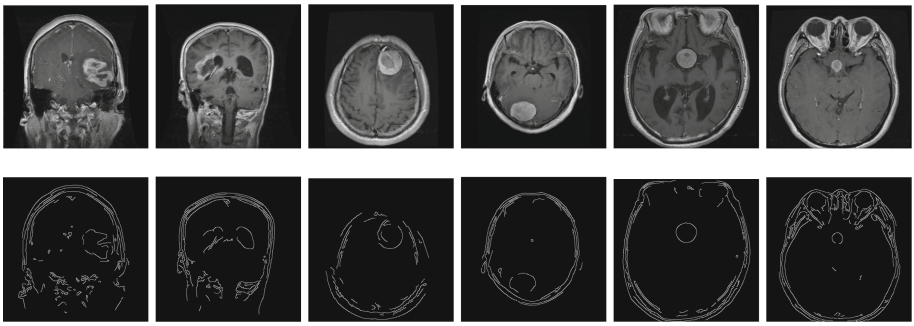


Fig. 1. Examples of MRI images and their corresponding edge images from the brain tumor dataset [29].

Canny edge detector to extract contours and edges from the original MRI image, the proposed architecture of our hierarchical classification model designed to differentiate between tumor types, and the model parameters set to optimize the model performance.

3.1 Data Processing with Canny Edge Detector

In optimizing MRI image processing for brain tumor detection, the Canny edge detector was chosen for its ability to accurately delineate tumor boundaries while ignoring irrelevant pixels, enhancing tumor identification precision [30]. The decision to utilize the Canny Edge Detector before model training considers the distinct morphologies of pituitary, glioma, and meningioma tumors. Pituitary tumors are typically situated between the eyes, gliomas, and meningiomas can occur in various brain parts, but meningiomas are generally round in shape. Thus, edge-enhanced images are advantageous for classification, isolating tumors from other MRI components like tissues and fluids, and streamlining model training.

This preference is rooted in the Canny detector's sophisticated, multi-stage algorithm, which combines noise reduction and precise edge detection, setting it apart from alternatives like Sobel, Prewitt, or Roberts cross [31]. This precision is crucial for MRI scans, where accurately defining tumor boundaries directly impacts diagnosis.

Our implementation of the Canny edge detector begins with converting MRI images to grayscale, emphasizing their structural integrity. The process continues with Gaussian blurring using a 5×5 kernel, which effectively reduces noise by smoothing pixel values with their neighbors based on a Gaussian distribution [32]. This step is crucial for minimizing potential false edge detection.

Next, we apply the Balance Contrast Enhancement Technique (BCET) to scale pixel values from 0 to 255, enhancing the overall image contrast and making critical features more distinguishable. Following this, K-means clustering with four clusters segments the MRI images into components: skull, brain tissues, fluid, and tumor, effectively removing unnecessary pixels and streamlining the application of the Canny edge detector. The Canny edge detector, applied with

Table 1. Parameters used in the image processing pipeline.

Step	Parameter	Value
Noise Reduction	Kernel Size	5
Contrast Enhancement	Low	0
	High	255
Segmentation	Clusters	4
Morphological Operations	Opening/Closing Structure	3×3
Edge Detection	Lower bound of the gradient	100
	Upper bound of the gradient	200

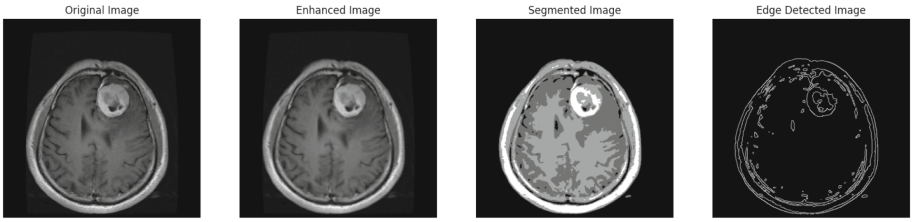


Fig. 2. The different stages in the image processing.

thresholds of 100 and 200, classifies edges based on gradient intensity: edges with intensities above 200 are marked as strong edges, and those between 100 and 200 are marked as weak edges. This method ensures that only the most distinct and relevant edges are preserved in the image. The algorithm calculates the intensity gradient at each pixel using operators like Sobel [33] to pinpoint potential edges through sharp intensity transitions. Non-maximum suppression then refines these edges, preserving only the most significant gradient pixels [34].

Based on their gradient magnitudes, double thresholding categorizes edges into strong and weak, followed by edge tracking by hysteresis. This final step solidifies the detection of meaningful edges by retaining weak edges only when connected to strong ones, effectively delineating tumor boundaries with remarkable precision [31]. This process is visualized in Fig. 1, where the edge images highlight the tumor regions in brain scans regardless of the image quality or orientation. The results of the processing pipeline are illustrated in Fig. 2, showcasing the original image, the enhanced image, the segmented image, and the final edge-detected image. The parameters of each method are listed in Table 1.

3.2 Model Architecture

In tackling the challenge of classifying brain tumors from MRI images, a detailed understanding of the visual traits of different tumor types is essential. Glioma and meningioma, two common brain tumor categories, present considerable similarities in their 2D MRI image appearances, especially in shape and position [9]. For more details, please refer to Fig. 4 and Fig. 5.

Automated classification models often struggle with the overlapping morphological and textural characteristics of brain tumor images, leading to confusion and inaccuracies. To address this, our model introduces a meta-classifier that first differentiates between visually distinct tumor types. Specifically, it separates the combined category of meningioma and glioma images from pituitary images. This initial broad differentiation helps streamline and enhance subsequent classification stages' accuracy.

Subsequently, a more specialized classifier is employed to distinguish between images resembling glioma and meningioma. This hierarchical strategy allows the model to efficiently categorize tumor images at a general level before honing in on discerning between glioma and meningioma, significantly lowering misclassification rates and improving diagnostic precision. Our approach involves this cascading strategy comprising primary and secondary binary classifiers to enhance classification precision through a systematic two-phase analytical approach (see Fig. 3).

For our primary classification, we selected ResNet50. ResNet50 [14] excels among neural network architectures due to its ability to learn deep features without encountering the vanishing gradient problem, a frequent issue in deep network training. We chose ResNet50 for classifying MRI brain tumors using the Canny edge detector because of its superior performance in initial experiments. Although we tested other models like GoogleNet, VGG16, VGG19, and InceptionV3, ResNet50 consistently achieved higher accuracy. Its deep architecture and residual learning capabilities make it adept at identifying crucial features within edge-enhanced images, which is essential for precise tumor detection.

In the hierarchical design of our model depicted in Fig. 3, the primary classifier distinguishes between pituitary and non-pituitary tumors, setting the stage for focused and accurate downstream classification. This foundational step is crucial, as it paves the way for subsequent, more granular classifications by ensuring that the preliminary groupings are precise and informative. Following this stage, images identified as non-pituitary advance to the secondary model, where a binary classifier dedicates its processing to differentiating between meningioma and glioma. Given their visual and textural similarities in MRI scans, this critical distinction between two often conflated tumor types underscores the necessity of a tailored and detailed approach.

Implementing a binary classifier as our secondary model was a strategic response to the challenge of differentiating between glioma and meningioma tumors in MRI images, which often display similar visual attributes. Our initial approach with a broad three-class model, utilizing a ResNet50 classifier, revealed significant overlap between these tumor types, leading to a high misclassification rate among the glioma and meningioma images.

Our approach underscores the need for a focused analysis to accurately identify subtle differences between glioma and meningioma, necessitating a binary classifier for this task. This specialized secondary model refines the primary classifier's broader effort by concentrating on distinguishing these two closely related tumor categories. A specially trained ResNet50-type network was considered for

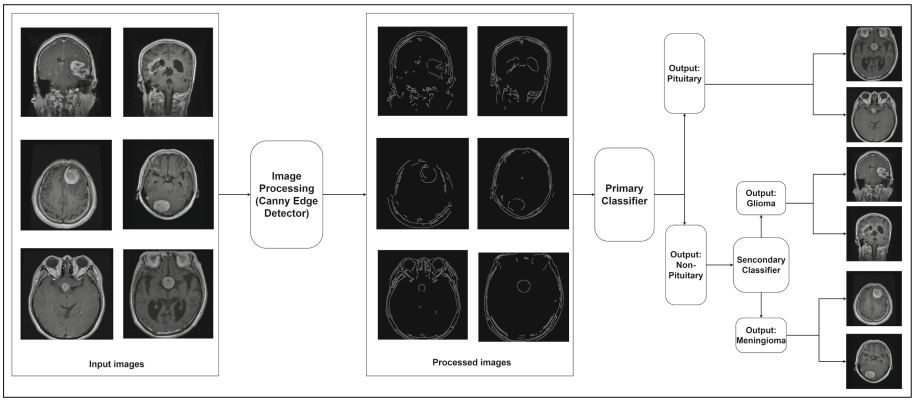


Fig. 3. System overview of the hierarchical classification model.

this model. The effectiveness of this approach is validated by our testing results, where the binary classifier significantly reduced misclassifications between these tumor types, enhancing model reliability and clinical diagnostic accuracy (see Sect. 4.5).

4 Experiments

Our study explores how our deep learning model handles the challenging brain tumor dataset by utilizing the Canny edge detector and the ResNet50 architecture. We also provide details in the upcoming sections about the evaluation metrics, computational setup, and hardware-related information considered for the experiments. The results are summarized in Sect. 4.5 followed by some comparison results with other state-of-the-art results obtained on the same benchmark data.

4.1 Data Description

The current research employs the brain tumor dataset provided by Cheng [29] available on Figshare[1], a well-known data collection used as a benchmark for many current research endeavors. This dataset comprises 3064 T1-weighted contrast-enhanced images from 233 patients labeled into three tumor categories: Meningioma, glioma, and pituitary. Some representative examples of these pathologies are to be seen in Fig 4, Fig. 5, and Fig. 6, respectively.

The dataset's inherent class imbalance poses a significant challenge, reflecting real-world diagnostic scenarios (see Table 2). For this study, the dataset was divided into a training set (70%) and a testing set (30%), ensuring a representative distribution of each tumor type in both subsets. This particular split for

[1] https://figshare.com/articles/dataset/brain_tumor_dataset/1512427.

Table 2. Number of MRI slices in the brain tumor dataset [29].

Tumor Class	Number of Patients	Number of MRI slices
Meningioma	82	708
Glioma	91	1426
Pituitary	60	930
Total	**233**	**3064**

the data was motivated by other researchers (see Table 5) who used the same split. Thus, we wanted to be able to compare our results directly with theirs and create a proper and fair evaluation framework.

4.2 Evaluation Metrics

To comprehensively assess the performance of our hierarchical brain tumor classification models, we utilize several key evaluation metrics [35,36]. Accuracy is used to measure the proportion of correct predictions out of the total predictions. Precision is defined as the ratio of true positives to the sum of true positives and false positives, reflecting the model's ability to correctly identify positive cases. Recall, or sensitivity, measures the ratio of true positives to the sum of true positives and false negatives, indicating the model's effectiveness in detecting all relevant cases. The F1 Score, which is the harmonic mean of precision and recall, provides a balanced measure of a model's performance, especially in situations where precision and recall may be imbalanced.

4.3 Training Parameters

The strategic application of data augmentation is crucial for enhancing the model's adaptability to new, unseen medical images, ensuring consistent performance across various imaging conditions. Using Keras's ImageDataGenerator class [37], we employ real-time augmentation techniques such as shearing, zooming, and flipping during training.

Data augmentation significantly increases the volume of training data. The original set of 2,144 training images (256×256 pixels) is expanded sixfold to 12,864 images post-augmentation, reducing overfitting and enhancing model accuracy in diagnosing new images.

We selected the Adam optimizer and categorical cross-entropy loss function [38] for model compilation, following best practices for multi-class classification in deep learning. The training was configured over 60 epochs, with a learning rate of 0.0001 and a batch size of 32. This setup balances precise model tuning and computational efficiency, optimized through several trial runs.

4.4 Computational Setup

Our experiments utilized Google Colab's cloud-based platform, leveraging the Tesla V100 GPU for its rapid processing capabilities, which provided necessary

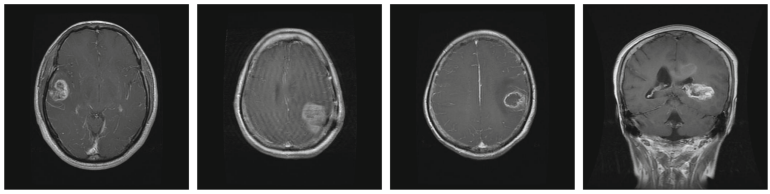


Fig. 4. Image examples from the glioma class.

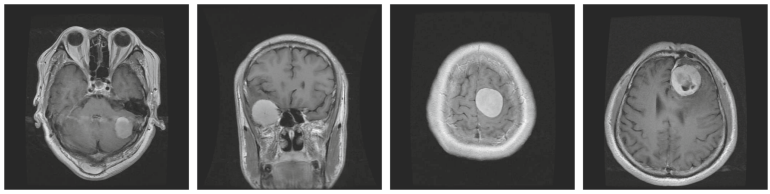


Fig. 5. Image examples from the meningioma class.

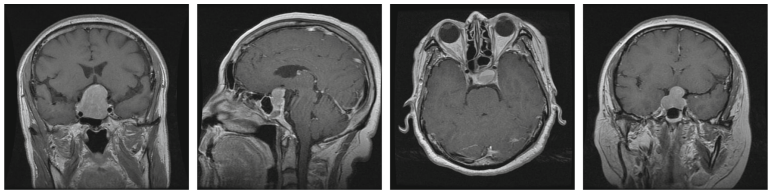


Fig. 6. Image examples from the pituitary class.

resources for managing the ResNet50 architecture and volume of data from datasets. TensorFlow [39] was employed for deep learning model construction, with Keras [40] providing an accessible interface for neural network design.

For image processing and augmentation, we used OpenCV [41], while NumPy [42] enabled efficient numerical computations. Matplotlib [43] and Seaborn [44] were used for data visualization, generating insightful plots and graphics to illustrate our models' results and performance metrics.

4.5 Results

After testing on 920 images (Glioma: 428, Meningioma: 213, Pituitary: 279) across five different randomly selected data splits, maintaining a 70% vs. 30% ratio, our hierarchical classification model achieved an exceptional average accuracy of 96.54%. The class-wise average accuracies were 98.2% for glioma, 93.4% for meningioma, and 96.3% for pituitary. Detailed evaluations for the five random runs with different datasets are shown in Table 3, including the exact class distributions. The relatively low accuracy for meningioma is attributed to its similarities with glioma, as reflected in the confusion matrices in Fig. 8.

Table 3. Results of the five test sets generated for the experiments.

Experiment	Glioma	Meningioma	Pituitary	Model Accuracy (%)
1	421	197	269	96.4
2	424	205	266	97.3
3	420	204	275	97.7
4	425	190	265	95.7
5	412	199	268	95.6
Average accuracy (%)	98.2	93.4	96.3	96.54

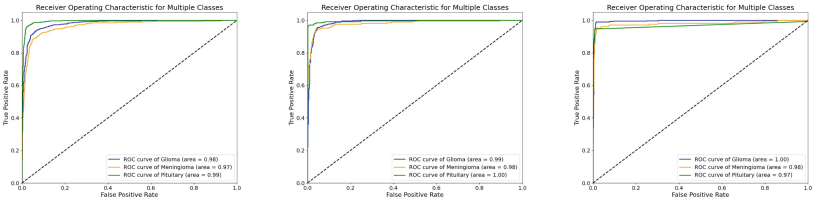
Table 4. Classification scores for the experiment 1.

Class	Precision	Recall	F1-score	Support
Glioma	0.95	0.98	0.97	428
Meningioma	0.96	0.92	0.94	213
Pituitary	0.99	0.96	0.98	279
Average Accuracy	96.41%			
Macro Avg	0.97	0.96	0.96	920
Weighted Avg	0.96	0.96	0.96	920

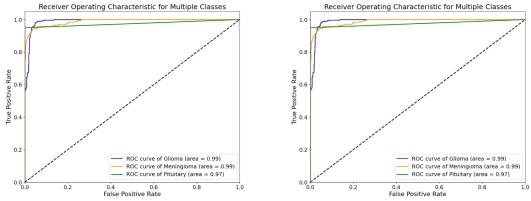
The classification report, detailed in Table 4, highlights the model's high precision, recall, and F1-scores across the three tumor types: glioma, meningioma, and pituitary-attesting to its robustness and reliability in medical diagnostics. The confusion matrices detailed in Fig. 8 show the model's proficiency in distinguishing pituitary tumors. However, the classification of glioma and meningioma is more difficult due to their similar imaging characteristics. To better qualify the results besides the classical accuracy measure, Fig. 7 separately shows the receiver operating characteristic (ROC) curve for all five experiments.

Initially, the model achieved 89% accuracy scores when a single classifier (see Fig. 8) was considered to distinguish between the three tumor types. However, introducing this two-stage hierarchical approach, which first isolates the pituitary class, significantly improved the overall performance, mitigating confusion between the glioma and meningioma classes by the second specialized network classifier.

Our model's efficiency is further highlighted by the computational speed, with the Canny edge detector processing features in just 0.00079 s per image, allowing for rapid classification of the test suite. Such speed, at an average of 0.225 s per image for classification over 920 images, underscores the model's potential for real-time clinical application without compromising accuracy. The "Macro Avg" and "Weighted Avg" metrics provide insight into the model's consistent performance across classes and its sensitivity to class imbalance, with a slight decrease in the weighted metrics indicating the impact of prevalence in the overall model's performance.

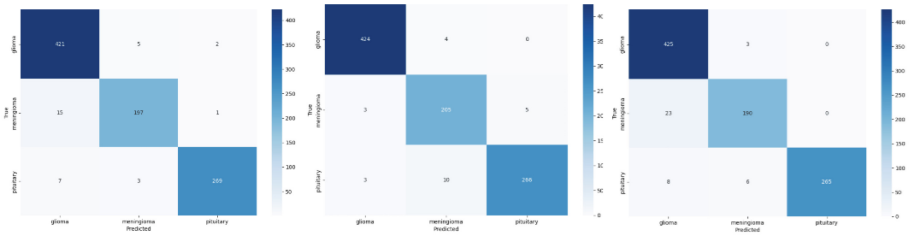


(a) ROC curve for experiment 1. (b) ROC curve for experiment 2. (c) ROC curve of experiment 3.

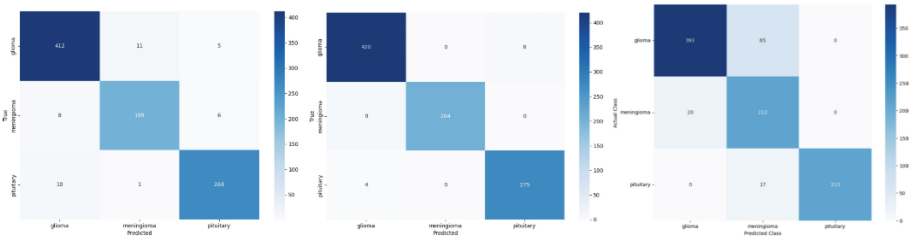


(d) ROC curve for experiment 4. (e) ROC curve for experiment 5.

Fig. 7. ROC curve of all five experiments based on Table 3.



(a) Confusion matrix of experiment 1. (b) Confusion matrix of experiment 2. (c) Confusion matrix of experiment 3.



(d) Confusion matrix of experiment 4. (e) Confusion matrix of experiment 5. (f) Confusion matrix of a single classifier.

Fig. 8. Confusion matrices for all experiments.

Table 5. Comparison of the proposed framework with the other state of art models.

Method	Classifier	Acc. (%)	Comment
Khan et al. [15]	23-layers CNN	**97.8**	Train: 2454 images Test: 610 images
Soumik et al. [16]	InceptionV3	**99.4**	5-fold cross-validation Train: 2452 images Test: 612 images
Swati et al. [17]	VGG19	94.5	5-fold cross-validation Train: 2452 images Test: 612 images
Irmak [18]	CNN	92.66	Train:2424 images Test: 640 images
Zhou et al. [21]	DenseNet-LSTM	93	Train: 2207 images Test: 857 images
Montoya et al. [22]	Resnet50	**97.3**	Train: 2451 images Test: 613 images
Dheepak et al. [23]	Distinct Customised Kernel (Ensemble) with SVM Classifier	93	5-fold cross-validation Train: 2452 images Test: 612 images
Cheng et al. [24]	SVM and KNN	91.2	5-fold cross-validation Train: 2452 images Test: 612 images
Afshar et al. [27]	Capsnet	90.8	5-fold cross-validation Train: 2452 images Test: 612 images
Shanaka et al. [28]	Deep Learning + Active Contouring	94.6	Train: 2144 images Test: 920 images
Momina et al. [45]	Mask RCNN + ResNet50	95.9	Train: 2144 images Test: 920 images
Proposed Method	Canny Edge Detector + ResNet50	**96.54**	Train: 2144 images Test: 920 images

4.6 Comparison with Other Methods

Table 5 encapsulates a comprehensive comparison, wherein each method, including our proposed approach, has been rigorously tested and benchmarked on the same brain tumor dataset comprising 3064 images using mainly similar splits for training and testing. This ensures a consistent and fair evaluation platform for all techniques under consideration, allowing for objectively assessing their relative performance in brain tumor classification for the data collection.

As reported in Table 5, some results [15,16,22] exceed ours, while the large majority of the works rank behind our achievements. For those works outperforming our strategy, after a thorough analysis, one could realize that these methods considered as input the original images in their original size, which puts an enormous computational burden on the models, or they use way more training data to fine-tune their models while the testing is performed only on a limited number of images which is way bellow to our 920 images considered for test.

5 Conclusion

In conclusion, this study introduces a novel methodology that synergizes Canny edge detection with a hierarchical deep learning classification scheme to significantly advance the accuracy of brain tumor classification from MRI scans. By harnessing the precision of edge detection techniques to enhance the visibility of tumor boundaries, coupled with the power of a two-tiered deep learning framework, we present an approach that markedly improves diagnostic processes in neuro-oncology.

Our methodology not only elevates classification accuracy to an impressive 96.54% rate but also delineates a path towards reducing the reliance on extensive human expertise in the initial stages of diagnosis.

The integration of the ResNet50 architecture, augmented through transfer learning, enables robust feature extraction and classification from edge images, effectively bypassing the significant variability inherent in the original MRI images. The results achieved, along with comparisons to other state-of-the-art methods, demonstrate the relevance and strength of our approach, positioning it among the top strategies in the literature. Future work will focus on further refining our model by exploring additional deep learning architectures and edge detection techniques, aiming to enhance both the accuracy and efficiency of tumor classification.

References

1. Moallem, G., et al.: Detecting and segmenting overlapping red blood cells in microscopic images of thin blood smears. In: Tomaszewski, J.E., Gurcan, M.N., (eds.) Medical Imaging 2018: Digital Pathology, Houston. SPIE Proceedings, vol. 10581, p. 105811F. Texas, United States, 10-15 February 2018 (2018)
2. Nkouanga, H.Y., Vajda, S.: Automatic tuberculosis detection using chest x-ray analysis with position enhanced structural information. In: 25th International Conference on Pattern Recognition, ICPR 2020, Virtual Event / Milan, Italy, January 10-15, 2021, pp. 6439–6446. IEEE (2020)
3. Vajda, S., et al.: Feature selection for automatic tuberculosis screening in frontal chest radiographs. J. Med. Syst. **42**(8):146:1–146:11 (2018)
4. Kaifi, R.: A review of recent advances in brain tumor diagnosis based on AI-based classification. Diagnostics, **13**(18) (2023)
5. Abdusalomov, A.B., Mukhiddinov, M., Whangbo, T.K.: Brain tumor detection based on deep learning approaches and magnetic resonance imaging. Cancers, **15**(16) (2023)

6. Ilic, I., Ilić, M.: International patterns and trends in the brain cancer incidence and mortality: an observational study based on the global burden of disease. Heliyon **9**, e18222 (2023)
7. Soumick, C., Faraz, N., Nürnberger, A., Oliver, S.: Classification of brain tumours in MR images using deep spatiospatial models. Sci. Rep. **12**(1) (2022)
8. Ranjbarzadeh, R., Caputo, A., Tirkolaee, E.B., Ghoushchi, S.J., Bendechache, M.: Brain tumor segmentation of MRI images: a comprehensive review on the application of artificial intelligence tools. Comput. Biol. Med. **152**, 106405 (2023)
9. Ce, M., et al.: Artificial intelligence in brain tumor imaging: a step toward personalized medicine. Current Oncol. **30**(3), 2673–2701 (2023)
10. Nazir, M., Shakil, S., Khurshid, K.: Role of deep learning in brain tumor detection and classification (2015 to 2020): a review. Comput. Med. Imaging Graph. **91**, 101940 (2021)
11. Sarfarazi, S., Toygar, Ö.: Classification of brain tumors on MRI images using deep learning architectures. In: 9th International IFS Contemporary Mathematics and Engineering Conference Special Issue, pp. 1177–1186 (2023)
12. Krizhevsky, A., Sutskever, I., Hinton, G.E.: ImageNet classification with deep convolutional neural networks. In: Advances in Neural Information Processing Systems, pp. 1097–1105 (2012)
13. Liu, S., Deng, W.: Very deep convolutional neural network based image classification using small training sample size. In: 2015 3rd IAPR Asian Conference on Pattern Recognition (ACPR), pp. 730–734 (2015)
14. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 770–778 (2016)
15. Khan, M.S.I., et al.: Accurate brain tumor detection using deep convolutional neural network. Comput. Struct. Biotechnol. J. **20**(2022) (2020)
16. Israk, F., Soumik, M., Ali, M.: Brain tumor classification with inception network based deep learning model using transfer learning. In: 2020 IEEE Region 10 Symposium (TENSYMP), pp. 1018–1021 (2020)
17. Swati, Z.N., Zhao, Q., Kabir, M., et al.: Brain tumor classification for MR images using transfer learning and fine-tuning. Comput. Med. Imaging Graph. **75**, 34–46 (2019)
18. Irmak, E.: Multi-classification of brain tumor MRI images using deep convolutional neural network with fully optimized framework. Iran. J. Sci. Technol. Trans. Electr. Eng. **45**, 1015–1036 (2021)
19. Rasheed, Z., et al.: Brain tumor classification from MRI using image enhancement and convolutional neural network techniques, vol. 13 (2023)
20. Akshya, K.S., Priyadarsan, P., Muralibabu, K.: Effective use of clustering techniques for brain tumor segmentation. In: 2023 IEEE 3rd International Conference on Applied Electromagnetics, Signal Processing, & Communication (AESPC), pp. 1–4 (2023)
21. Zhou, Y., et al.: Holistic brain tumor screening and classification based on DenseNet and recurrent neural network. In: Crimi, A., Bakas, S., Kuijf, H., Keyvan, F., Reyes, M., van Walsum, T. (eds.) BrainLes 2018. LNCS, vol. 11383, pp. 208–217. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-11723-8_21
22. Montoya, S.F.A., Rojas, A.E., Vásquez, L.F.N.: Classification of brain tumors: a comparative approach of shallow and deep neural networks. SN Comput. Sci. **5**(142) (2024)
23. Dheepak, G., Christaline, J.A., Vaishali, D.: Mehw-svm multi-kernel approach for improved brain tumour classification. IET Image Process. (2023)

24. Cheng, J., et al.: Enhanced performance of brain tumor classification via tumor region augmentation and partition. PLOS ONE **10**(10), 1–13 (2015)
25. Raza, A., et al.: A hybrid deep learning-based approach for brain tumor classification. Electronics, **11**(7) (2022)
26. Ullah, M.S., Attique Khan, M., Masood, A., Mzoughi, O., Saidani, O., Alturki, N.: Brain tumor classification from MRI scans: a framework of hybrid deep learning model with Bayesian optimization and quantum theory-based marine predator algorithm. Front. Oncol. **14** (2024)
27. Afshar, P., Plataniotis, K.N., Mohammadi, A.: Capsule networks for brain tumor classification based on mri images and coarse tumor boundaries. In: ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 1368–1372 (2019)
28. Gunasekara, S.R., Kaldera, H.N.T.K., Dissanayake, M.B.: A systematic approach for MRI brain tumor localization and segmentation using deep learning and active contouring. J. Healthc. Eng. **1–13**, 2021 (2021)
29. Cheng, J.: Brain tumor dataset 2017 (2024)
30. Bharodiya, A., Gonsai, A.: An improved edge detection algorithm for x-ray images based on the statistical range. Heliyon **5**, e02743 (2019)
31. Canny, J.: A computational approach to edge detection. IEEE Trans. Pattern Anal. Mach. Intell. PAMI **8**(6), 679–698 (1986)
32. Gedraite, E.S., Hadad, M.: Investigation on the effect of a gaussian blur in image filtering and segmentation. In: Proceedings ELMAR-2011, pp. 393–396 (2011)
33. Katiyar, S.K., Arun, P.V.: Comparative analysis of common edge detection techniques in context of object extraction. IEEE Trans. Geosci. Remote Sens. **50**(11), 68–79 (2012)
34. Acton, S.T.: Chapter 20 - diffusion partial differential equations for edge detection. In: Bovik, A., (ed.) The Essential Guide to Image Processing, pp. 525–552. Academic Press, Boston (2009)
35. Vujovic, Z.D.: Classification model evaluation metrics. Int. J. Adv. Comput. Sci. Appl. (IJACSA) **12**(6) (2021)
36. Fawcett, T.: An introduction to roc analysis. Pattern Recogn. Lett. 27(8), 861–874 (2006). ROC Analysis in Pattern Recognition
37. TensorFlow. tf.keras.preprocessing.image.ImageDataGenerator documentation (2023). Accessed 24 Mar 2024
38. Kingma, D.P., Ba, J.: Adam: a method for stochastic optimization (2017)
39. Abadi, M., Agarwal, A., et al.: TensorFlow: Large-scale machine learning on heterogeneous systems. Software available from tensorflow.org
40. Chollet, F et al.: Keras (2015)
41. Bradski, G.: The OpenCV library. Dr. Dobb's J. Softw. Tools (2000)
42. Harris, C.R., et al.: Array programming with NumPy. Nature **585**(7825), 357–362 (2020)
43. Hunter, J.D.: Matplotlib: A 2D graphics environment. Comput. Sci. Eng. **9**(3), 90–95 (2007)
44. Waskom, M.L.: seaborn: statistical data visualization. J. Open Source Softw. **6**(60), 3021 (2021)
45. Masood, M., et al.: A novel deep learning method for recognition and classification of brain tumors from MRI images. Diagnostics, **11**(5) (2021)

Shape Induced Multi-class Deep Graph Cut for Hippocampus Subfield Segmentation

Arijit De and Ananda S. Chowdhury(✉)

Jadavpur University, Kolkata 700032, West Bengal, India
{arijitde.etce.rs,as.chowdhury}@jadavpuruniversity.in

Abstract. Automated evaluation of hippocampus volume plays a crucial role in the analysis of various neurodegenerative conditions like Alzheimer's Disease and Epilepsy. Examination of the hippocampus subfields assumes paramount importance as it can reveal early signs of brain abnormalities. However, delineating these subfields becomes extremely challenging due to their intricate nature and the requirement for manually annotated high-resolution magnetic resonance images. In this paper, we propose an innovative deep graph cut approach, boosted by shape information, for automatic segmentation of hippocampus subfields. A deep learned shape term is incorporated in the energy function of the graph cut. A modified $\alpha - \beta$ swap technique, that leverages deep learning, is designed to improve the execution time of the proposed multi-class segmentation algorithm. We demonstrate the efficacy of our solution by outperforming a number of state-of-the-art methods on the publicly available Kulaga-Yoskovitz dataset.

Keywords: Hippocampus subfield segmentation · Multi-class Graph cut · Shape term · Deep Learning

1 Introduction

The hippocampus (HC) is a paired brain structure situated in the medial temporal lobe adjacent to the brainstem in close proximity to the cerebellum. It plays a crucial role in various cognitive functions, such as, memory and spatial reasoning [23]. Over the past decade, there has been a growing interest in segmenting hippocampal subfields using MRI. Recent research has identified distinct functional roles for these anatomical subregions, with CA1 implicated in memory integration and inference [28], CA3 in memory retrieval [8], and both the dentate gyrus (DG) and CA3 in pattern separation [2]. Clinically, the volume or morphology of the hippocampus and its subfields are closely related to many neurodegenerative diseases like Epilepsy [32] and Alzheimer's disease [17]. So, it is desirable to develop automatic hippocampal subfields segmentation from brain MR image. However, manual delineation of hippocampal subfields is a laborious and time-intensive task, leading to constraints on sample sizes in various studies.

A. Antonacopoulos et al. (Eds.): ICPR 2024, LNCS 15313, pp. 242–257, 2025.
https://doi.org/10.1007/978-3-031-78201-5_16

We first discuss some works that use classical techniques for HC subfield segmentation. The authors in [36] used a multi-atlas approach combined with a similarity-weighted voting and a boosting-based error correction as a solution. They termed their method, ASHS. This method took several hours to produce a segmentation due to exhaustive use of non-linear registrations. More recently, a method named HIPS [26] obtained state-of-the-art results with relatively low processing times. While classical methods integrated domain-specific image features like gradient, intensity, and textures within an energy minimization framework, they are found to heavily depend on initialization, such as manual seeding. As a result, they are prone to segmentation errors caused by uncertain positioning of the markers. Furthermore, these approaches are quite laborious and may not be practical for clinical environments with a heavy workload.

Deep learning approaches have surpassed classical methods, delivering superior segmentation performance in significantly less time. Recently, due to the expansion of deep learning (DL) in medical imaging for tasks like classification [10] and segmentation [11], novel methods based on this technology have been proposed to further improve the accuracy of HC sub-field segmentation. UNet based methods [22,37] have shown promising results. Shi et al. [29] proposed a Generative Adversarial Network (GAN) to create a segmentation model. But UNet and GAN based methods require a lot of data and medical imaging lacks consistent and sufficient annotated data, making DL algorithms perform poorly in many cases [31]. Although some authors have tried to bypass this problem using multi scaling technique [35] and using higher resolution data [20], the problem continues to exist.

A combination of both classical and DL techniques can achieve better segmentation performance than using DL methods or classical methods in isolation. For example, see the works [24,27] in lung nodule segmentation. In case of 3D brain tumor segmentation [12], a combination of UNet and graph cut helped circumvent manual seeding problem of the graph cut and undersegmentation of UNet due to scarcity of data.

In this paper, we propose a shape driven multi-class segmentation method using UNet [9] and graph cut. We take inspirations from [12], [11] and [33] to create a state-of-the-art model to segment the HC into three classes, namely, CA1-3, CA4/DG and Subiculum (Sub). In [5], Boykov et al. showed that a two-class segmentation is achievable in polynomial time using graph cuts. However, if the number of labels exceeds 2 (as for the present problem), finding an exact solution becomes an NP-hard problem. They suggested two types of large moves (changing labels of individual pixels/voxels) based on minimal graph cuts, namely, α-expansion and α-β swap. In this work, we use UNet to improve the α-β swap. A number of research papers demonstrated that use of shape information can improve the segmentation accuracy [11,19,21]. Shape priors provide valuable guidance by incorporating prior knowledge about the expected shapes of objects in the image. This guidance helps the segmentation algorithm to make more informed decisions about the boundaries and regions of interest [30]. By imposing shape constraints, shape priors help to enforce consistency in the segmented

shapes, ensuring that the output conforms to the expected shape characteristics [1]. The shape prior has to be made adaptive in case of substantial noise and intensity variations. Here, we better an adaptive shape prior from deep learned information via UNet. Our main contributions are now summarized below:

1. Propose a new energy function for multi-class segmentation based on graph cut and deep learning (UNet)
2. Incorporate learned information from UNet for optimizing the number of α-β swaps
3. Show how an adaptive shape prior can be learned from UNet

2 Proposed Method

2.1 Deep Graph Cut

Let us define the 3D MRI input as a gray-scale volumetric data, which may be represented as a 3D weighted graph denoted by $G = G(V, E)$. Each vertex is represented by a voxel x in G, and X is the collection of all voxels. We introduce two new vertices, called 'source' and 'sink', represented by s and t respectively. There are two sorts of edges or linkages that we consider: t-links (T) and n-links (N). s and t is linked to very voxel x through t-links. We utilize a compact 26-neighborhood, represented as $Ne(x)$ for every voxel x. Assume that y is a neighbor of x. Therefore, y belongs to the neighborhood of x, and we establish a connection between x and y by an n-link. Therefore, the set V is defined as the union of sets X, s, and t, whereas the set E is defined as the union of sets T and N. Let us establish a segmentation A as a classification of all voxels into two distinct classes: "object" or "background". This classification is done on a voxel-wise basis. Therefore, according to the reference [7], it is necessary to minimize the subsequent energy function:

$$\zeta(A) = B(A) + \lambda R(A) \tag{1}$$

The term $B(A)$ represents the boundary characteristics or smoothness term of A, while $R(A)$ represents the regional properties or data term of A. These terms are represented mathematically as below:

$$B(A) = \sum_{x \in X, y \in Ne(x)} B_{(x,y)} \tag{2}$$

$$R(A) = \sum_{x \in X} R_x \tag{3}$$

In [12], we modified the above energy function (Eq. 1) by incorporating learned information from the 3D UNet [9]. The modified energy function is given below:

$$\begin{aligned} \zeta_{DGC}(A) = & \sum_{x \in X, y \in Ne(x)} B_{DGC}(x, y) + \\ & \lambda_{DGC}(x) \sum_{x \in X} R_{DGC}(x) \end{aligned} \tag{4}$$

2.2 Multi-class Deep Graph Cut

As stated earlier, in this work, we deal with multi-class hippocampus segmentation where a voxel x can belong to any of 'CA1-3', 'CA4/DG' and Subiculum. Following [5], our goal is to find a labeling f that assigns each voxel $x \in X$ a label $f_x \in \mathcal{L}$ and, $|\mathcal{L}| > 2$, where f is both piecewise smooth and consistent with the observed data. Any labeling f can be uniquely represented by a partition of image voxels, $V = V_l | l \in \mathcal{L}$ where $V_l = x \in V | f_x = l$ is a subset of pixels assigned a label l. Hence, Eq. 4 can be rewritten as:

$$\zeta_{DGC}(A_f) = \sum_{x \in X, y \in Ne(x)} B_{DGC}(f_x, f_y) + \lambda_{DGC}(f_x) \sum_{x \in X} R_{DGC}(f_x) \tag{5}$$

where $\zeta_{DGC}(A_f)$ is the energy of the labelling f. The knowledge acquired from the 3D UNet [9] is included into the energy function of the 3D graph cut algorithm in order to achieve precise segmentation. The 3D probability map is obtained from the last convolutional layer for each image. This map is then used to determine the probability, denoted as $Pr(f_x)_{UN}$, of any voxel x belonging to label f_x. The 3D UNet calculates a regression function that maps the voxels of a 3D input to a 3D voxel-wise probability map. This is denoted as $\mathcal{P} : \mathbb{R}^3 \rightarrow (0, 1)$, and it assigns a value between 0 and 1 to each voxel. Additionally, this probability map is used as an automated seed required by 3D graph cut algorithm. With this, we now explain the smoothness and data term in the context of multi label problem as follows. As mentioned in [12], $B_{DGC}(f_x, f_y)$ is a product of four components as shown below-

$$B_{DGC}(f_x, f_y) = K_{(x,y)} \times e^{-(\frac{(I_x - I_y)^2}{2\sigma^2})} \times \frac{1}{d(x,y)} \times \frac{1}{\delta(x,y)_{DGC}} \tag{6}$$

where $d(x, y)$ represents the Euclidean distance between two voxels x and y having intensity values I_x and I_y respectively. The term $K_{(x,y)}$ is based on the probabilities of x and y to have the labeling f_x and f_y and is mathematically represented as:

$$K_{(x,y)} = 1 - |Pr(f_x)_{UN} - Pr(f_y)_{UN}| \tag{7}$$

where $f_x = f_y$. The factor σ is the standard deviation of voxel intensities of the image [19]. The term $\delta(x, y)_{DGC}$ denotes the sum of differences between probabilities of neighbouring voxels x and y to belong to f_x and f_y where $f_x \neq f_y$. This can be expressed as:

$$\begin{aligned} \delta(x,y)_{DGC} = &|Pr(f_x = \alpha)_{UN} - Pr(f_y = \alpha)_{UN}| \\ &+ |Pr(f_x = \beta)_{UN} - Pr(f_y = \beta)_{UN}| \end{aligned} \tag{8}$$

where $\alpha, \beta \in \mathcal{L}$. The data term $R_{DGC}(f_x)$ is dependent on the probability map of UNet as shown below-

$$R_{DGC}(f_x) = -\ln Pr(f_x = \alpha)_{UN} \tag{9}$$

α-β Swap. As mentioned in Boykov et al.'s article [5], segmenting a binary image is possible in polynomial time using graph cut. But if the number of labels is more than 2 (as in our case), finding exact solution becomes an NP-hard problem. Therefore, Boykov et al. proposed two types of moves based on minimal graph cuts - α-expansion move and α-β swap move. A standard move means changing the label of a single vertex (voxel in our case). We now discuss how and why we modify the optimization function of α-β swap moves. The choice of which type of move to select depends on whether the smoothness term of the energy function is a metric or a semi-metric [5]. If the smoothness term is metric, α-expansion can be used, otherwise α-β swap move needs to be used. For a function $V(\alpha, \beta)$ to be metric, it has to satisfy the following constraints:

1. $V(\alpha, \beta) \Leftrightarrow \alpha = \beta$
2. $V(\alpha, \beta) = V(\beta, \alpha) \geq 0$
3. $V(\alpha, \beta) \leq V(\alpha, \gamma) + V(\gamma, \alpha)$

for any labels $\alpha, \beta, \gamma \in \mathcal{L}$ [5]. If $V(\alpha, \beta)$ satisfies only the constraints (1) and (2) but not (3), then it is called a semi-metric. We have chosen α-β swap moves to optimize our energy function, as our smoothness term ($B_{DGC}(f_x, f_y)$) is a semi-metric. We explicitly show in the appendix that the smoothness term is indeed a semi-metric.

Deep Learned α-β Swap. If a move from a partition V_l to a new partition V_l' has labels α, β, then $V_l' = V_l$ for all labels $l \neq \alpha, \beta$. This is known as a α-β swap [5]. So, the only thing that's different between V_l and V_l' is that some voxels that were labeled as α in V_l are now labeled as β, and the other way around. The main idea is to use graph cuts to separate all α voxels from β voxels one by one. Each time through the algorithm, the $\alpha - \beta$ mix will be different. The program will keep going through all the possible combinations until it converges with the minimum energy. The algorithm is guaranteed to converge in $O(V)$ time, but when there are a lot of vertices, the whole segmentation process takes a long time.

As reported in [5], segmenting a 384×288 image with $\alpha - \beta$ swap takes 35 seconds. In our case, the image size is $182 \times 218 \times 182$ which is far greater than the images used in [5]. Hence, there is a dire need to optimise the move algorithm to speed up the overall segmentation.

For this, we turn to Pseudo-Boolean optimization techniques used in [6]. As mentioned in [6], we encode the moves of the $\alpha - \beta$ swap algorithm as a vector of binary variables $t = t_i, \forall_i \in V$. So, $t_i = 0$ means the label of voxel i changed to α and $t_i = 1$ means the label changed to β. The transformation function $T(f^c, t)$ of a move algorithm takes the current labelling f^c and a move t and returns a new labelling f^n that has been induced by the move. The transformation function $T_{\alpha\beta}()$ for an $\alpha - \beta$ swap transforms f^c as

$$f_i^n = T_{\alpha\beta}(f_i^c, t_i) = \begin{cases} f_i^c, & \text{if } f_i^c \neq \alpha \text{ and } f_i^c \neq \beta, \\ \alpha, & \text{if } f_i^c = \alpha \text{ or } \beta \text{ and } t_i = 0, \\ \beta, & \text{if } f_i^c = \alpha \text{ or } \beta \text{ and } t_i = 1. \end{cases} \tag{10}$$

If the current labelling f_i^c is neither α nor β, we don't change it. The energy of the move t is the energy of labelling f^n that the move t induces, i.e., $E_m(t0 = E(T(f^c mt))$. Further details about the pseudo boolean energy of the swap move can be found in Sect. 3.3 of [4].

We modify Eq. 10 by adding the label probability information derived from UNet as follows-

$$f_i^n = T(f_i^c, t_i) = \begin{cases} f_i^c, & \text{if } f_i^c \neq \alpha \text{ and } f_i^c \neq \beta, \\ & \\ \alpha, & \text{if } f_i^c = \alpha \text{ or } \beta, t_i = 0, [Pr(f_x = \alpha) - Pr(f_x = \beta) > \tau] \\ & , \text{ or } [Pr(f_x = \beta) - Pr(f_x = \alpha) < (1 - \tau)], \\ & \\ \beta, & \text{if } f_i^c = \alpha \text{ or } \beta, t_i = 1, [Pr(f_x = \beta) - Pr(f_x = \alpha) > \tau] \\ & , \text{ or } [Pr(f_x = \alpha) - Pr(f_x = \beta) < (1 - \tau)], \end{cases} \tag{11}$$

We added two more constraints when deciding the new labelling to be α or β. We wanted the confidence of the UNet model to decide whether a label should be swapped or should be kept the same. We define confidence of prediction as the difference between the probabilities of a voxel to have label α and β, i.e. $[Pr(f_x = \alpha) - Pr(f_x = \beta)]$. Generally, a model is said to predict a label (say α) with high confidence if the probability of the voxel to belong α is much higher than that of the voxel to belong to another label (say β), i.e., if $[Pr(f_x = \alpha) - Pr(f_x = \beta)] > \tau$ or $[Pr(f_x = \beta) - Pr(f_x = \alpha) < (1 - \tau)]$ where τ is some threshold. So, if for any voxel i, $[Pr(f_x = \alpha) - Pr(f_x = \beta)] < \tau$ and f_i^c, the current label of i is either α or β, then the label of voxel i will be changed to α. Similarly, the decision to change a label to β if the confidence of the UNet model for that voxel to have label β is greater than τ.

2.3 Deep Learned Shape Information

In this section, we first discuss what is the significance of addition of a shape term in HC segmentation, then we briefly mention the importance of adaptive shape term and how the incorporation of UNet's probability map helps in creating an adaptive shape term suitable for 3D HC segmentation.

Need of an Adaptive Shape Term. In cases where images are affected by substantial noise and intensity variations, the necessity for a shape prior can vary across different pixels. Consequently, assigning a uniform weight to the shape prior term for all pixels may not be suitable. In our case, 3D MRI images do suffer from noise and intensity variations and in many places the HC and non HC region of the brain has very low contrast as shown in Fig. 1. Segmentation tasks use the adaptive shape term to selectively impose shape constraints based on pixel labeling difficulty to give flexibility and local adaptation. This adaptability allows the system to modify shape prior strength based on local image properties,

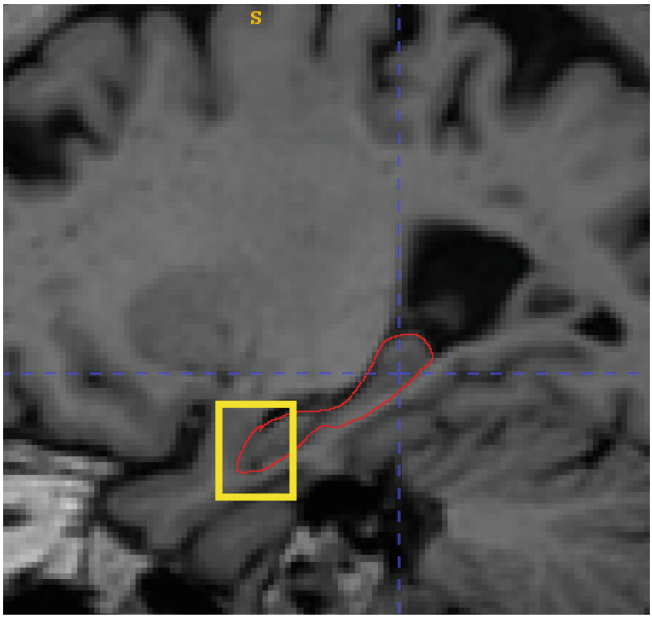


Fig. 1. A sagittal slice view of a brain MRI showing the hippocampus bounded in red. The region marked in yellow shows that the contrast is less between hippocampus and its surrounding region which poses a challenge in segmentation. (Color figure online)

applying shape restrictions where they are most useful. The adaptive shape term dynamically adjusts shape priors based on image intensity, resulting in more accurate and context-aware segmentation results [33].

Improved Adaptive Shape Term with UNet. We improve Wang et al.'s [33] adaptive shape prior formulation using learned information from 3D UNet. Following their approach, we add to the smoothness term $B_{DGC}(f_x, f_y)$, a shape term of the form $S_{DGC}(f_x, f_y)$ with η as the shape weight. Note that the authors in [33] defined $\eta = e^{-(Pr(x)-Pr(y))^2}$, where, $Pr(k)$ is the likelihood of a pixel k belonging to the foreground. They determined this likelihood by using an unsupervised technique like applying Gaussian filter.

Unlike in [33], where the authors used 2D images and performed binary segmentation, we deal with 3D images and multi-class segmentation in this work. So we redefine η as:

$$\eta = e^{-(Pr(f_x)-Pr(f_y))^2} \tag{12}$$

where, $Pr(f_k)$ denotes the likelihood of voxel k to have labelling f_k. Further, $Pr(f_k)$ is obtained from the probability map of UNet as mentioned in Sect. 2.1. This ensures that we have better probability values than that obtained from using unsupervised techniques, as in [33].

The shape term, S_{DGC}, can be formulated as the unsigned distance function (as used in [16]) of the segmentation obtained after thresholding probability map $\mathcal{P}$. Let the segmentation obtained by thresholding $\mathcal{P}$ is $\mathcal{G}$ with a threshold value of κ. Then,

$$S_{DGC} = \bar{\phi}_{\mathcal{G}} \left(\frac{x+y}{2} \right) \tag{13}$$

where, $\bar{\phi}_{\mathcal{G}} : \mathfrak{R}^3 \rightarrow \mathfrak{R}$ is the distance function on $\mathcal{G}$ and is such that $\bar{c} = x \in \mathfrak{R}^3 : \bar{\phi}(x) = 0$; $\bar{c}$ being the set of points that form the boundary of the shape. The energy will be low if $\bar{\phi}_{\mathcal{G}} \left(\frac{x+y}{2} \right) \approx 0$ for all neighboring voxels x and y and $f_x \neq f_y$. If a voxel x lies near the shape template, then it will satisfy $\bar{\phi}(x) \approx 0$. Since, $\left(\frac{x+y}{2} \right)$ is roughly a point on the boundary of the segmented object, the condition for S_{DGC} to be small is the same as the condition that the boundary of the segmented object lies near the shape template.

2.4 Shape Driven Multi-class Deep Graph Cut

We started with Eq. 4 which is the deep graph cut for energy function for binary segmentation. Then we modified it to adapt to multi class segmentation in Eq. 14. We then modified the α-β swap moves using information from UNet as described in Sect. 2.2. We compute the data term $R_{DGC}(f_x)$, smoothness term $B'_{DGC}(f_x, f_y)$ and Finally, after incorporating the two terms described in the previous section and shown in Eq. 12 and Eq. 13 in the energy function of multi class deep graph cut (Eq. 5), we get the final energy function for Shape induced Multi class Deep Graph Cut (SMDGC) method as shown below-

$$\begin{aligned} \zeta_{SMDGC} = & \sum_{x \in X, y \in Ne(x), f_x \neq f_y} B_{DGC}(f_x, f_y) + \eta S_{DGC}(f_x, f_y) + \\ & \lambda_{DGC}\left(f_x\right) \sum_{x \in X} R_{DGC}\left(f_x\right) \end{aligned} \tag{14}$$

The algorithm for our overall workflow is shown below in Algorithm 1 followed by a discussion on its time complexity-

2.5 Analysis of Time-Complexity

The alpha-beta swap algorithm, used for multi-label graph cuts, iteratively optimizes the graph $G = G(V, E)$ by swapping labels between pairs (α, β) to minimize the energy function. The time-complexity of this algorithm is influenced by factors, such as, the number of labels, the number of pixels (or nodes), and the underlying max-flow algorithm used. The time-complexity analysis of the proposed SMDGC algorithm is as follows:

1. Max-flow Computation: The time-complexity of each max-flow computation depends on the specific max-flow algorithm used. We have used Edmond Karp's approach [14] in the Ford-Fulkerson algorithm [15], where augmenting paths are computed using the Breadth First Search. It has a time-complexity of $O(VE^2)$.

Algorithm 1: SMDGC

Input: 3D UNet model M trained on training set of the data, Graph G represented as a 3D grid of voxels
Output: G_{out} with desired segmentation and same dimensions as G

1 Compute data term $R_{DGC}(f_x)$ for each voxel in $x \in G$ as shown in Eq. 9
2 Compute smoothness term $B_{DGC}(f_x, f_y)$ for each voxel x and its neighbor $y \in N_e(x)$ in G as shown in Eq. 6
3 Compute η and S_{DGC} as shown in Eqs. 12 and 13 respectively
4 Compute the modified transformation function for optimizing the number of α-β swap moves using Eq. 11.
5 Compute the final energy function using Eq. 14, perform Graph cut and store the result in G_{out}.
6 **return** G_{out}

2. Number of Labels: The alpha-beta swap considers all pairs of labels. So, the number of iterations is proportional to $\binom{\mathcal{L}}{2} = \frac{\mathcal{L}(\mathcal{L}-1)}{2}$, which is $O(\mathcal{L}^2)$, where $\mathcal{L}$ denotes the number of labels.

Combining the two factors, we can say that the time-complexity of our algorithm is $O(\mathcal{L}^2 V E^2)$.

3 Experimental Results

In this section, we first describe the dataset used for the experimentation. Necessary details of the parameters, and, hyperparameters used, are provided next. To showcase the significance of different components of our solution, we then present a number of ablation studies. Finally, we show performance comparisons with several state-of-the-art approaches.

3.1 Dataset

We have used a publicly available 3D MRI dataset, described in [18]. We henceforth abbreviate this Kulaga-Yoskovitz dataset as the KY dataset. It comprises 25 healthy adult subjects aged between 21 and 53 years, with a mean age of 31.2 ± 7.5 years and a male-to-female ratio of 12 : 13. The data were acquired using a 3T Siemens Tim Trio MRI scanner equipped with a 32-channel head coil. Submillimeter T1 and T2 images were obtained for all participants. The 3D MPRAGE T1 image had a spatial resolution of $0.6 \times 0.6 \times 0.6$ mm^3 (isotropic voxel size). The matrix size was 336×384, with a field of view (FOV) of 201 mm $\times$ 229 mm and 240 axial slices at a slice thickness of 0.6 mm. The T2 image was acquired using a 2D turbo spin-echo sequence with a matrix size of 512×512, an FOV of 203 mm $\times$ 203 mm, and 60 coronal slices angled perpendicular to the hippocampal long axis, with a slice thickness of 2 mm, resulting in a voxel size of $0.4 \times 0.4 \times 2.0$ mm^3. The manual segmentation protocol for this dataset categorized the hippocampus into three labels: subiculum (SUB), a combination of CA1, CA2, and CA3 (CA1-3), and a combination of CA4 and DG (CA4/DG).

3.2 Preprocessing

We preprocessed the data using the steps described in [22] that included cropping along the HC area and data augmentation by left - right flipping. Finally, we got 50 samples having 100 axial slices with the length and breadth same as mentioned earlier for T1 and T2 images. Since, the number of samples is less, we did 5-fold cross validation to assess the results.

3.3 Experimental Settings

We implemented our 3D UNet network in PyTorch [25]. The training process was executed on a HP-Z640 workstation having Intel Xeon processor with 14 Cores, a Random Access Memory (RAM) with capacity of 128 GB along with a dedicated graphics processor unit (GPU) of 24 GB with model name NVIDIA Titan RTX. The network is trained for 100 epochs with initial learning rate of 0.0001, weight decay of 0.00001 and mini-batch size equal to 2 samples. We have used Adam Optimizer and dice loss in the process. Our energy function does not have any parameters that need to be set manually, as all the information is being provided by a trained UNet model. The only parameter κ, which is used for thresholding the UNet probabilities to create a segmentation (as described in Sect. 2.3) is set to, 0.5 which is the most common value as mentioned in [34].

We have used Dice score [13] as the metric to compare the segmentation performance, as this was the metric used by most other works on segmentation [20,29,37].

3.4 Ablation Studies

As mentioned earlier, we present three ablation studies for providing a better understanding of our solution. Table 1 shows the results of our first ablation study. Here, we demonstrate the improvement our model brings over baseline multi-class graph cut and multi-class UNet, applied in isolation. In Table 2, we first show the performance of DGC [12] for multi-class segmentation, by adding to it traditional α-β swap. It is then demonstrated how the execution time improves due to optimization of α-β swap strategy using learned information from UNet. The computation time improves drastically improves by almost 50%

Table 1. Ablation Study I: Comparison of segmentation performance of multi-class graph cut, multi-class UNet, and the proposed method. Mean Dice Score of each competing approach over all three classes are reported. Best values are shown in **bold**.

Method	Dice Score
Multi-class Graph Cut	0.64000 ± 0.073
Multi-class UNet	0.82000 ± 0.047
SMDGC (Ours)	**0.91467 ± 0.009**

Table 2. Ablation Study II: Impact of deep learned α-β swap on the segmentation performance. Mean Dice Score of each competing approach over all three classes are reported. Best values are shown in **bold**.

Method	Dice Score	Time (in secs)
DGC [12] with α-β swap	0.88230 $\pm$ 0.032	15
DGC [12] with deep learned α-β swap	**0.89860 $\pm$ 0.008**	**8**

Table 3. Ablation Study III: Impact of deep learned shape on segmentation performance. Mean Dice Score of each competing approach over all three classes are reported. Best values are shown in **bold**.

Method	Dice Score
Graph cut with adaptive shape term [33]	0.85297 $\pm$ 0.026
SMDGC (Ours)	**0.91467 $\pm$ 0.009**

when we use modified α-β swap. DGC with normal α-β swap takes on average 15 s to segment one 3D sample, whereas, the modified α-β swap achieves the same goal in 8 s. A slight improvement in the segmentation accuracy can also be noticed as more informed decision is taken to change a label during a move due to the confidence of UNet incorporated into the optimization strategy (Eq. 11).

We then analyze the impact of a shape term in a graph cut setup through Table 3. The first row shows the results, where an adaptive shape term is used but without deep learning. For that, we re-implement [33] and add α-β swap moves, as that work was originally developed for binary segmentation. We compare the performance of this method with ours, where we have employed deep learned shape information (Eq. 14). The values of the Dice Scores clearly illustrate the benefits of a deep learned shape information.

Qualitative comparisons of different strategies used in the three ablation studies are shown in Fig. 2. We only include the second method of Table 2 as the improvement there is more in terms of execution time to achieve desired segmentation, rather than the segmentation accuracy per se. In Fig. 2, the visual improvements in segmentation performances clearly corroborate the quantitative results. We specifically highlight how multi-class UNet and graph cut with adaptive shape term suffers from over segmentation of CA1-3 and SUB, as shown in the yellow boxes of the sagittal slices. DGC with modified α-β swap also finds it difficult to decide among the CA1-3 and SUB, as shown in the yellow box of its coronal slice. In general, CA4/DG is relatively difficult to segment by all methods, as it is the smallest region among the three classes under consideration.

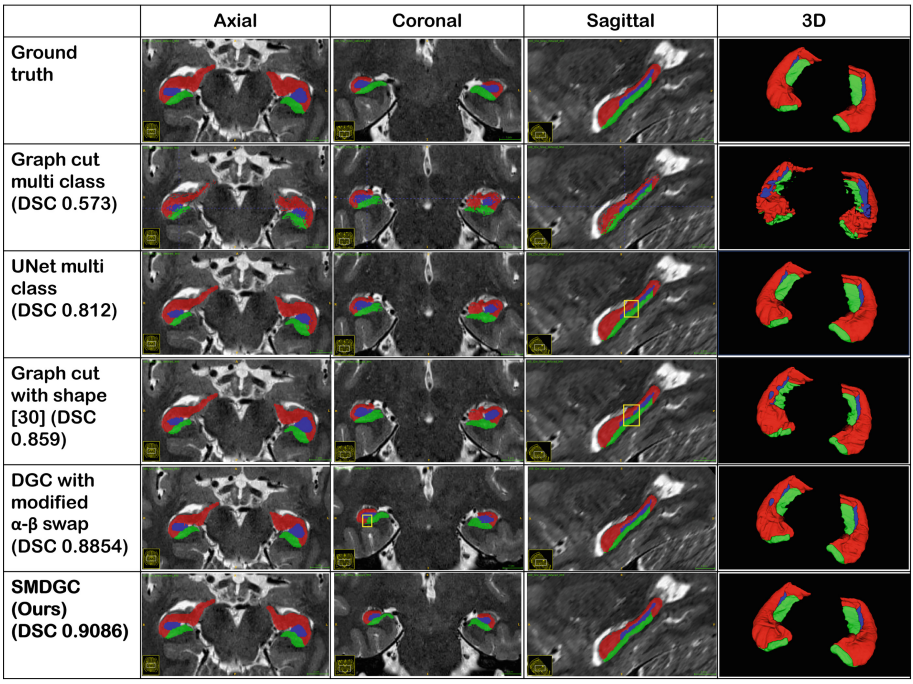


Fig. 2. Qualitative ablation of our method. GT represents the ground truth. Segmentation with color red represents CA1-3 class, blue represents the CA4/DG class and green represents the SUB class. (Color figure online)

3.5 Comparison with State-of-the-Art Methods

We compare our proposed method with five state-of-the-art approaches (papers published within the last five years). These methods are [20,22,35,37], and, [29]. We showed comparisons with only DL based approaches, as we did not come across any work on multi-class HC segmentation using primarily graph cuts. As can be clearly seen from the Table 4, our method has yielded the highest mean Dice Score, which is marginally better than [29]. We are marginally behind [29] in the SUB subfield segmentation, the most complex object to segment within the HC. However, our model requires much less computational resource, as we used only plain 3D UNet, and, graph cut while, other approaches have used sophisticated DL models that take a lot of time and resources to train.

Table 4. Comparison with state-of-the-art methods. Mean Dice Score ± standard deviation of Dice Score is reported for each class. Additionally, the overall mean Dice Score is reported for each method. Best values are shown in **bold**.

Method	CA1-3	CA4/DG	SUB	Mean
Syn SegNet (2023) [20]	0.865 ± 0.005	0.821 ± 0.014	0.821 ± 0.013	0.835 ± 0.007
CAST (2020) [35]	0.917 ± 0.011	0.89 ± 0.017	0.881 ± 0.021	0.906 ± 0.014
ResDUNet (2019) [37]	0.92 ± 0.011	0.879 ± 0.02	0.888 ± 0.018	0.896
UNet CNN (2022) [22]	0.9245 ± 0.01	0.8887 ± 0.023	0.898 ± 0.015	0.9
GANs (2019) [29]	0.919	0.903	**0.906**	0.88
SMDGC (Ours)	**0.933 ± 0.007**	**0.9078 ± 0.013**	0.903 ± 0.008	**0.9146 ± 0.009**

4 Conclusion

Hippocampus subfield segmentation is a crucial step in the diagnosis of many diseases like Alzheimer's, Epilepsy as the treatment depends on the analysis of volumetric atrophy of the subfields. Automating this process will greatly enhance the treatment experience for both doctors and patients. In this work, we proposed a state-of-the-art method of subfield segmentation using a combination of multi-class graph cuts with shape information and deep learning. In particular, we showed how deep learning can boost the shape knowledge, and the $\alpha - \beta$ swap move. Comparisons with a number of state-of-the-art methods on a publicly available dataset clearly establish the efficacy of our proposed solution. In the future, we plan to include other datasets that contain more subfields to achieve a more fine-grained segmentation of the hippocampus. We also plan to introduce explainability [3] into our proposed hippocampus segmentation model so that it can be more effectively used in the real-world clinical settings.

Acknowledgement. Arijit De was supported by Tata Consultancy Services Research Scholar Program (TCS-RSP).

Appendix

Lemma 1. *The product of a semi-metric and a metric function is semi-metric.*

Proof. Let ρ_1 be a semi-metric function and ρ_2 be a metric function defined on some set X. Then, for any $x, y, z \in X$:

$$\begin{aligned} \rho_1(x,z) &\leq \rho_1(x,y) + \rho_1(y,z) \quad \left(\because \rho_1(x,z) \text{ is semi-metric}\right) \\ &= c_1 \cdot (\rho_1(x,y) + \rho_1(y,z)) \quad \left(\text{where } c_1 \leq 1\right) \end{aligned}$$

$$\begin{aligned} \rho_2(x,z) &> \rho_2(x,y) + \rho_2(y,z) \quad \left(\because \rho_2(x,z) \text{ is metric}\right) \\ &= c_2 \cdot (\rho_2(x,y) + \rho_2(y,z)) \quad \left(\text{where } c_2 > 1\right) \end{aligned}$$

Now, consider the product

$$\begin{aligned}\rho(x,z) &= \rho_1(x,z)\cdot\rho_2(x,z)\\ &= \Big(c_1\left(\rho_1(x,y)+\rho_1(y,z)\right)\Big)\cdot\Big(c_2\left(\rho_2(x,y)+\rho_2(y,z)\right)\Big)\\ &= c_1c_2\rho_1(x,y)\rho_2(x,y)+c_1c_2\rho_1(x,y)\rho_2(y,z)\\ &\quad + c_1c_2\rho_2(y,z)\rho_1(x,y)+c_1c_2\rho_2(y,z)\rho_1(y,z)\end{aligned}$$

Now, $c_1c_2 > 1$ when $c_1 = 1$. Hence,

$$\begin{aligned}\rho(x,z) &> \rho_1(x,y)\rho_2(x,y)+\rho_1(y,z)\rho_2(y,z)+\rho_1(x,y)\rho_2(y,z)+\rho_2(y,z)\rho_1(x,y)\\ &> \rho(x,y)+\rho(y,z)+\omega_1+\omega_2\end{aligned}$$

where $\omega_1 \geq 0$ and $\omega_2 \geq 0$.
Therefore, $\rho(x,z) > \rho(x,y)+\rho(y,z)$ when $c_1 = 1$ which means $\rho(x,z)$ does not obey triangle inequality for some particular cases. Thus, the product of a semi-metric and a metric function remains a semi-metric function.

Theorem 1. *$B_{DGC}(f_x, f_y)$ is a semi-metric.*

Proof. B_{DGC} is a product of four components as shown in Eq. 6. Among them, $K_{(x,y)}$ and $\delta(x,y)_{DGC}$ depends on probabilities. From Eq. 7, it is evident that $K_{(x,y)}$ lie between $[0,1]$ whereas, from Eq. 8, $\delta(x,y)_{DGC}$ lie between $[0,2]$. Both these functions satisfy points (1) and (2) of Sect. 2.2, i.e., $K_{x,y} \Leftrightarrow x = y$, $K_{x,y} = K_{y,x} \geq 0$ and similarly for $\delta(x,y)_{DGC}$. But they do not satisfy the triangle inequality (point 3). If we consider three voxels x, y and z, then $K_{x,y}$, $K_{y,z}$ and $K_{x,z}$ can take any value between $[0,1]$ and hence, there will be cases where $K_{(x,z)} > K_{(x,y)} + K_{(y,z)}$ for some x, y and z. A similar situation can also occur in the case of $\delta(x,y)_{DGC}$. Therefore, these functions are semi metric. For example, if we consider $K_{(x,y)} = 0.2$, $K_{(y,z)} = 0.3$ and $K_{(x,z)} = 0.7$, then $K_{(x,z)} > K_{(x,y)} + K_{(y,z)}$. Now, we consider the term $e^{-(\frac{(I_x - I_y)^2}{2\sigma^2})}$, which is based on image intensities I_x and I_y. The intensity value lies between $[0, 255]$. We can similarly argue that $e^{-(\frac{(I_x - I_z)^2}{2\sigma^2})} > e^{-(\frac{(I_x - I_y)^2}{2\sigma^2})} + e^{-(\frac{(I_y - I_z)^2}{2\sigma^2})}$ for some I_x, I_y and I_z.

Thus, $K_{(x,y)}$, $e^{-(\frac{(I_x - I_y)^2}{2\sigma^2})}$ and $\delta(x,y)_{DGC}$ are semi metric in nature, $\frac{1}{d(x,y)}$ is metric as $d(x,y)$ is the Euclidean distance. Hence, from Lemma 1 it follows that $B_{DGC}(x,y)$ is a semi metric.

References

1. Ali, A.M., Farag, A.A., El-Baz, A.S.: Graph cuts framework for kidney segmentation with prior shape constraints. In: Ayache, N., Ourselin, S., Maeder, A. (eds.) MICCAI 2007. LNCS, vol. 4791, pp. 384–392. Springer, Heidelberg (2007). https://doi.org/10.1007/978-3-540-75757-3_47
2. Baker, S., et al.: The human dentate gyrus plays a necessary role in discriminating new memories. Curr. Biol. **26**, 2629–2634 (2016)

3. Band, S.S., et al.: Application of explainable artificial intelligence in medical health: a systematic review of interpretability methods. Inform. Med. Unlocked **40**, 101286 (2023)
4. Blake, A., Kohli, P., Rother, C.: Markov Random Fields for Vision and Image Processing. MIT Press, Cambridge (2011)
5. Boykov, Y., Veksler, O., Zabih, R.: Fast approximate energy minimization via graph cuts. IEEE TPAMI **23**(11), 1222–1239 (2001)
6. Boykov, Y., Veksler, O., Zabih, R.: Optimizing multilabel mrfs using move-making algorithms. Markov Random Fields for Vision and Image Processing, pp. 51–64 (2011)
7. Boykov, Y.Y., Jolly, M.P.: Interactive graph cuts for optimal boundary & region segmentation of objects in nd images. In: ICCV, vol. 1, pp. 105–112. IEEE (2001)
8. Chadwick, M.J., Bonnici, H.M., Maguire, E.A.: CA3 size predicts the precision of memory recall. Proc. Natl. Acad. Sci. **111**(29), 10720–10725 (2014)
9. Çiçek, Ö., Abdulkadir, A., Lienkamp, S.S., Brox, T., Ronneberger, O.: 3D U-Net: learning dense volumetric segmentation from sparse annotation. In: Ourselin, S., Joskowicz, L., Sabuncu, M.R., Unal, G., Wells, W. (eds.) MICCAI 2016. LNCS, vol. 9901, pp. 424–432. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46723-8_49
10. De, A., Mhatre, R., Tiwari, M., Chowdhury, A.S.: Brain tumor classification from radiology and histopathology using deep features and graph convolutional network. In: 2022 26th International Conference on Pattern Recognition (ICPR), pp. 4420–4426 (2022)
11. De, A., Tiwari, M., Chowdhury, A.S.: 3D hippocampus segmentation using a hog based loss function with majority pooling. In: IEEE ICIP, pp. 2260–2264. IEEE (2023)
12. De, A., Tiwari, M., Grisan, E., Chowdhury, A.S.: A deep graph cut model for 3D brain tumor segmentation. In: EMBC, pp. 2105–2109. IEEE (2022)
13. Dice, L.R.: Measures of the amount of ecologic association between species. Ecology **26**(3), 297–302 (1945)
14. Edmonds, J., Karp, R.M.: Theoretical improvements in algorithmic efficiency for network flow problems. J. ACM (JACM) **19**(2), 248–264 (1972)
15. Ford, L., Fulkerson, D.: Flows in Networks. Princeton Landmarks in Mathematics and Physics, Princeton University Press (2015). https://books.google.co.in/books?id=fw7WCgAAQBAJ
16. Freedman, D., Zhang, T.: Interactive graph cut based segmentation with shape priors. In: IEEE CVPR, vol. 1, pp. 755–762. IEEE (2005)
17. Hobbs, K.H., Zhang, P., Shi, B., Smith, C.D., Liu, J.: Quad-mesh based radial distance biomarkers for Alzheimer's disease. In: Proceedings - International Symposium on Biomedical Imaging **2016-June**, pp. 19–23 (2016)
18. Kulaga-Yoskovitz, J., et al.: Multi-contrast submillimetric 3 Tesla hippocampal subfield segmentation protocol and dataset. Sci. Data **2**(1), 1–9 (2015)
19. Li, G., et al.: Automatic liver segmentation based on shape constraints and deformable graph cut in CT images. IEEE Trans. Image Process. **24**(12), 5315–5329 (2015)
20. Li, X., et al.: Syn_SegNet: a joint deep neural network for ultrahigh-field 7T MRI synthesis and hippocampal subfield segmentation in routine 3T MRI. IEEE J. Biomed. Health Inform. **27**(10), 4866–4877 (2023)
21. Malcolm, J., Rathi, Y., Tannenbaum, A.: Graph cut segmentation with nonlinear shape priors. In: IEEE ICIP, vol. 4, pp. IV–365. IEEE (2007)

22. Manjón, J.V., Romero, J.E., Coupe, P.: A novel deep learning based hippocampus subfield segmentation method. Sci. Rep. **12**(1) (2022)
23. Milner, B.: Psychological defects produced by temporal lobe excision. Res. Publ. Assoc. Res. Nerv. Ment. Dis. **36**, 244–257 (1958)
24. Mukherjee, S., Huang, X., Bhagalia, R.R.: Lung nodule segmentation using deep learned prior based graph cut. In: IEEE ISBI, pp. 1205–1208. IEEE (2017)
25. Paszke, A., et al.: PyTorch: an imperative style, high-performance deep learning library. In: Proceedings of the 33rd International Conference on Neural Information Processing Systems. Curran Associates Inc., Red Hook, NY, USA (2019)
26. Romero, J.E., Coupé, P., Manjón, J.V.: HIPS: a new hippocampus subfield segmentation method. Neuroimage **163**, 286–295 (2017)
27. Roy, R., Chakraborti, T., Chowdhury, A.S.: A deep learning-shape driven level set synergism for pulmonary nodule segmentation. Pattern Recogn. Lett. **123**, 31–38 (2019)
28. Schlichting, M.L., Zeithamova, D., Preston, A.R.: CA1 subfield contributions to memory integration and inference. Hippocampus **24**(10), 1248–1260 (2014)
29. Shi, Y., Cheng, K., Liu, Z.: Hippocampal subfields segmentation in brain MR images using generative adversarial networks. BioMed. Eng. Online **18**(1) (2019)
30. Slabaugh, G., Unal, G.: Graph cuts segmentation using an elliptical shape prior. In: IEEE ICIP, vol. 2, pp. II–1222. IEEE (2005)
31. Suganyadevi, S., Seethalakshmi, V., Balasamy, K.: A review on deep learning in medical image analysis. Int. J. Multimed. Inf. Retrieval **11**(1), 19–38 (2021)
32. Voets, N.L., Bernhardt, B.C., Kim, H., Yoon, U., Bernasconi, N.: Increased temporolimbic cortical folding complexity in temporal lobe epilepsy. Neurology. **76**(2), 138–144 (2010)
33. Wang, H., Zhang, H., Ray, N.: Adaptive shape prior in graph cut image segmentation. Pattern Recogn. **46**(5), 1409–1414 (2013)
34. Xu, Y., Gao, F., Wu, T., Bennett, K.M., Charlton, J.R., Sarkar, S.: U-net with optimal thresholding for small blob detection in medical images. In: 2019 IEEE 15th International Conference on Automation Science and Engineering (CASE), pp. 1761–1767 (2019)
35. Yang, Z., Zhuang, X., Mishra, V., Sreenivasan, K., Cordes, D.: CAST: a multi-scale convolutional neural network based automated hippocampal subfield segmentation toolbox. NeuroImage **218**, 116947 (2020)
36. Yushkevich, P.A., et al.: Automated volumetry and regional thickness analysis of hippocampal subfields and medial temporal cortical structures in mild cognitive impairment. Human Brain Mapping **36**(1), 258–287 (2015)
37. Zhu, H., et al.: Dilated dense U-net for infant hippocampus subfield segmentation. Front. Neuroinform. **13** (2019)

Tract-RLFormer: A Tract-Specific RL Policy Based Decoder-Only Transformer Network

Ankita Joshi[1(✉)], Ashutosh Sharma[1], Anoushkrit Goel[1], Ranjeet Ranjan Jha[2], Chirag Kamal Ahuja[3], Arnav Bhavsar[1], and Aditya Nigam[1]

[1] Indian Institute of Technology (IIT) Mandi, Mandi , India
s22041@students.iitmandi.ac.in
[2] Indian Institute of Technology (IIT) Patna, Patna, India
[3] Post-Graduate Institute of Medical Eduation and Research (PGIMER), Chandigarh, India

Abstract. Fiber tractography is a cornerstone of neuroimaging, enabling the detailed mapping of the brain's white matter pathways through diffusion MRI. This is crucial for understanding brain connectivity and function, making it a valuable tool in neurological applications. Despite its importance, tractography faces challenges due to its complexity and susceptibility to false positives, misrepresenting vital pathways. To address these issues, recent strategies have shifted towards deep learning, utilizing supervised learning, which depends on precise ground truth, or reinforcement learning, which operates without it. In this work, we propose Tract-RLFormer, a network utilizing both supervised and reinforcement learning, in a two-stage policy refinement process that markedly improves the accuracy and generalizability across various data-sets. By employing a tract-specific approach, our network directly delineates the tracts of interest, bypassing the traditional segmentation process. Through rigorous validation on datasets such as TractoInferno, HCP, and ISMRM-2015, our methodology demonstrates a leap forward in tractography, showcasing its ability to accurately map the brain's white matter tracts.

Keywords: Tractography · Transformers · Reinforcement Learning

1 Introduction

Tractography is an advanced reconstruction technique in neuroscience, that leverages diffusion MRI to create detailed visual representations of brain's white matter pathways. This technology has played a crucial role in assisting neurosurgeons with meticulous pre-surgical planning, benefiting patients with a range of neurological disorders [7], by enabling a deeper analysis of the white matter. Over the years, a range of tractography algorithms have been developed, to map critical neurological pathways. Deterministic algorithms [2] trace fiber paths directly based on the most probable direction of water molecule diffusion, offering clear but sometimes oversimplified views of white matter tracts. In contrast,

A. Antonacopoulos et al. (Eds.): ICPR 2024, LNCS 15313, pp. 258–275, 2025.
https://doi.org/10.1007/978-3-031-78201-5_17

probabilistic algorithms [4] incorporate the inherent uncertainty in diffusion data to predict multiple potential pathways, resulting in detailed fiber reconstruction. Global algorithms [8] attempt to reconcile the deterministic and probabilistic approaches by optimizing whole-brain tractography reconstructions to capture the complex architecture of brain connectivity.

Despite these advancements, tractography still faces challenges such as the *crossing-fibers* issue (also known as the *bottleneck* phenomenon) [12] due to its ill-posed nature. These issues arise because the algorithms rely on local diffusion information to reconstruct the brain's complete fiber network, occasionally resulting in erroneous projection of fiber pathways or false positive connections.

To overcome these obstacles, recent research employs machine learning and deep learning (DL) approaches to enhance tractography accuracy. Supervised DL techniques [3,15] for tractography rely on accurate and comprehensive ground truth data to train and validate the algorithms, which is very difficult to obtain. In this regard, recent works [20,21] have proposed deep reinforcement learning (DRL)-based approaches, that learn to perform tractography by interacting with the environment. These techniques, leveraging deep neural networks, enhance the ability to predict brain fiber configurations, promising significant advancements in fiber mapping quality for neurological research and clinical applications. However, improving tractography algorithms for effective use across diverse datasets remains a challenge for further research in the field.

Recently, transformers have shown remarkable performance in various domains, including language modeling [23], image recognition [6], time series forecasting [27], and even protein structure prediction [11]. Their robust performance in various sequence prediction tasks demonstrates their ability to capture long-range dependencies and contextual information effectively, making them well-suited for mapping neural pathways in tractography. Building on their success in related fields, we now extend the generalization, transfer learning, and autoregressive capabilities of transformers (GPT), to the tractography domain **in a novel hybrid framework**. We adopt an RL framework [5] to generate training data for our GPT model, Tract-RLFormer, reducing the need for extensive ground-truth typically needed to train transformers. This addresses a significant challenge of applying transformers for tractography where ground-truth fibers are very difficult to obtain.

This approach also represents a significant departure from traditional methods, as it simplifies the tractography process by targeting specific tracts, thereby eliminating the need for complex and often cumbersome segmentation algorithms employed post-tractography. Moreover our data-driven approach has the potential to utilize data from RL agents trained across diverse neuroimaging environments. Our key contributions are as follows:

1. **Data driven policy learning via hybrid framework**: We propose Tract-RLFormer, a GPT-based network trained by leveraging both reinforcement learning (RL) and supervised learning (SL) paradigms, to approximate and refine a policy for tract generation that outperforms recent RL-algorithms.

2. **Innovative Tract-Specific Generation:** We train Tract-RLFormer to generate the tract of interest, utilizing our developed Mask Refinement Module (MRM) to generate tracking masks for the target tract, bypassing segmentation overhead.
3. **Generalization:** Through extensive testing on diverse datasets (TractoInferno, HCP, ISMRM2015), we demonstrate our network's superior performance and generalization capabilities across different neuroimaging contexts.

2 Related Work

Research in fiber tractography has transitioned from traditional deterministic and probabilistic methods to machine learning and deep reinforcement learning. Supervised learning and the exploratory dynamics of deep reinforcement learning unlock several possibilities for accurately mapping the brain's connectivity. Below, we review some recent works in these paradigms.

Supervised Machine Learning Based Algorithms: In several studies, machine learning techniques have been explored to enhance fiber tractography with promising results. Notably, [13,14] utilized a Random Forest classifier in a supervised learning setting to identify 25 distinct fiber bundles, leveraging data from the ISMRM2015 dataset [12]. The effectiveness of their approach was assessed using the Tractometer tool, demonstrating the classifier's ability to accurately distinguish between different fiber pathways. Building on this foundation, subsequent research shifted focus towards regression-based methods for fiber tracking. [15] suggested to employ a Gated-Recurrent Unit (GRU) model to predict new tracking steps from diffusion signal resampled to 100 directions. This method advances the field, moving beyond traditional classification techniques to offer a more nuanced understanding of fiber tract development. Advancing the understanding of deep learning's potential for tractography, [3] applied both deterministic and probabilistic approaches to the task. In [25], authors introduced an innovative method known as iFOD3, utilizing a feed-forward neural network to analyze raw, resampled signals. This approach considers the spatial context of streamlines, incorporating seed points located at the interface between white and gray matter—a notable departure from conventional methods that focus solely on white matter. This broader perspective on seed point placement contributed to the method's enhanced performance. In a subsequent development in [24], authors presented a probabilistic machine learning model that outputs Fischer-von-Mises distributions rather than deterministic paths. This approach marked improvement over previous techniques, offering a more accurate and effective means of mapping the intricate networks of brain fiber tracts. These advancements underscore the rapidly evolving landscape of fiber tractography, highlighting the critical role of machine learning and deep learning in pushing the boundaries of neuroimaging research.

Reinforcement Learning Based Algorithms: Contrary to the supervised training common in machine and deep learning approaches (poses challenges due to the difficulty of generating large scale ground truth data), the authors

explored reinforcement learning (RL)-based approach for fiber tractography in [21]. In this approach, tractography is conducted similar to classical methods, wherein a reward function is employed by a learning model to generate streamlines based on local fiber orientation. Unlike the supervised paradigm for tractography, the RL-based model does not utilize reference streamlines while training. In [21], the Twin-Delayed Deep-Deterministic Policy Gradient (TD3) algorithm [9] and the Soft Actor-Critic (SAC) algorithm were employed for RL-based fiber tractography to reduce false positives and enhance model generalization. In [20], the authors further examined different aspects of the RL framework, such as algorithm choice, seeding strategies, state representation, and reward functions, paving the way for advancements in this domain.

3 Proposed Methodology

In this section, we begin by discussing the data and its preprocessing, followed by a systematic presentation of our proposal. We utilize three public diffusion MRI datasets (Table 1). These datasets include a series of diffusion weighted images (DWI) that capture the diffusion of water molecules in tissue. Each voxel in a DWI contains information about the magnitude and direction of water diffusion, reflecting the underlying tissue micro-structure.

Table 1. Description of the three public DWI Datasets

Dataset	Subjects	DWI data	Distortion Corrections
TractoInferno [16]	284	b = 1000 s/mm^2; resolution = 1 mm isometric	N4 bias field; eddy-current; head-motion
HCP [22]	1200	b = 1000/2000/3000 s/mm^2; 270 directions; resolution = 1.25 mm isometric	EPI; eddy-current; subject-motion
ISMRM [12]	1	b = 1000 s/mm^2; 32 directions; resolution = 2 mm isometric	eddy currents; head motion *(by our preprocessing)*

Diffusion MRI Pre-processing: We process DWI data to extract crucial information, including Spherical Harmonics Coefficients (SHC), Fiber Orientation Distribution Functions (fODF), and fiber peaks. Initially, the DWI data is projected into an 8^{th} order spherical harmonics basis, yielding 45 SHC volumes. The fODF, representing the distribution of fiber orientations within each voxel, is then computed, providing essential local information regarding streamline orientation. Subsequently, using the fODF, local fiber directions (peaks) are computed, which are used to define the reward function for training networks in the RL framework, as elaborated in Sect. 3.2.

Moreover, in traditional tractography methods, white matter masks are typically derived from DWI data to perform whole-brain tractography, followed by

segmentation of specific tracts. In contrast, we generate tailored masks for each tract, as detailed in Sect. 3.1. Our models are trained and tested within these masks, allowing for precise and efficient tract-specific analysis.

We propose an *iterative policy learning framework for tract-specific generation*, delineated as a five-step process (see Fig. 1). In this framework, we start by training an RL agent (TD3) to learn a policy by exploration (within the tracking mask) to generate a tract of interest. We call it as level-1 policy. Using this initial policy, the agent interacts with the (tracking) environment by taking actions (tracking steps). The agent's experience (policy rollouts) is collected and sampled to train a refined version of the policy, by our T-RLF model, which learns in a data-driven manner through general pre-training and tract-specific fine-tuning. Our study focuses on seven principal white matter (WM) tracts: Corpus Callosum (CC), left and right Pyramidal (PYT), Arcuate Fasciculus (AF), and Cingulum (CG) Tracts. The selection of these seven tracts is based on their clinical significance and frequent analysis as suggested in [18,20]. To conduct such tract-specific training and generation, we first compute a tracking region of interest (mask) tailored for each tract using our Mask Refinement Module (MRM), described in 3.1. Following this, we proceed with the five sequential steps depicted in Fig. 1, detailed in subsequent subsections of the methodology.

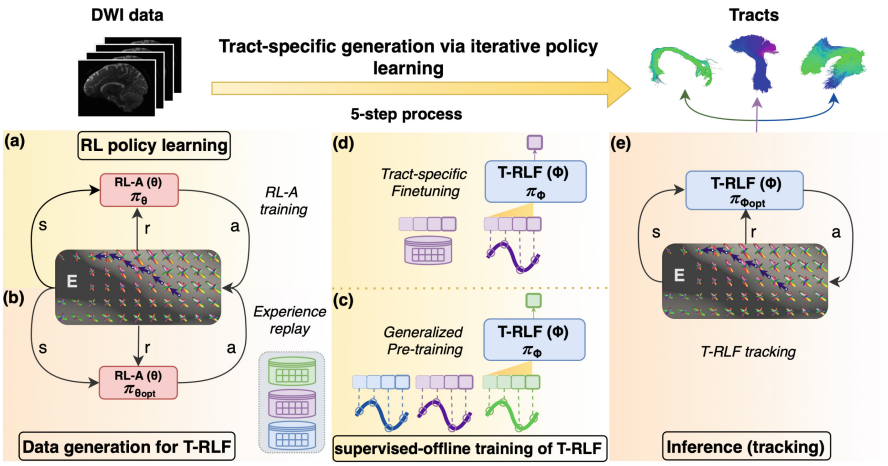


Fig. 1. Overview of the proposed Iterative Policy Learning for Tract-Specific Generation using DWI data. (a) An RL agent (π_θ) interacts with the environment (E) to learn an optimal **level-1 policy** ($\pi_{\theta opt}$). (b) This policy is used to generate tract-specific roll-outs, denoted as 'experience replay'. (c) and (d) illustrate the offline, auto-regressive training of the proposed Tract-RLFormer ϕ, referred to as T-RLF, over these roll-outs. In (c), T-RLF undergoes general pre-training, while in (d) it is fine-tuned to learn an optimal tract-specific policy ($\pi_{\phi opt}$). (e) shows the testing phase, where T-RLF, which has learned the new **level-2 policy** ($\pi_{\phi_{opt}}$), performs tracking in environment E to produce the desired tract. Training and tracking steps are shown in yellow and orange backgrounds, respectively.

3.1 Mask Refinement Module (MRM)

We combine reference tracts from 2 Atlases, namely HCP842 [26], and RecobundlesX [17] to develop a fiber template for each of the seven tract classes. To obtain the mask of a given tract for any subject, the template fibers of the tract are aligned to the subject's brain space [1], creating an initial mask which is then dilated by 5 mm to get an augmented region of interest (ROI). This ROI is further refined by our Mask Refinement Module (MRM), which produces a tracking mask for a specific tract utilizing the fiber orientation information of the given subject. It consists of a fully connected neural network (FCNN) that refines the augumented ROI to obtain an estimate of the ground-truth mask for a given subject. The process starts with a larger mask and progressively refines it by eliminating its voxels based on the Spherical Harmonics Coefficients (SHC) in the local neighborhood. The input for each voxel is the SHC (45 per voxel) of the voxel itself and its six immediate neighbors, concatenated with the expanded mask values, resulting in an input size of 322 (7 * 46).

The neural **network architecture** comprises three hidden layers with 512, 256, and 128 neurons, respectively. Each layer employs a ReLU activation function and is followed by batch normalization and a dropout layer (0.5). The output layer uses a sigmoid activation function, which determines the probability of retaining each voxel in the refined mask. Voxels with an output probability greater than 0.5 are kept in the predicted mask, while those with lower probabilities are eliminated. **Training** is performed voxel-wise, using Binary Cross Entropy as the loss function to compare the predicted mask value with the ground truth for each voxel. The model was trained with 50 subjects randomly selected from the TractoInferno dataset over 100 epochs. The resulting mask is then dilated by 1 mm to produce the final refined tracking mask for the given subject.

3.2 RL Policy Learning

We learn a Level-1 policy by training a reinforcement learning (RL) agent (π_θ) to perform fiber tracking. The RL agent learns the policy through exploration within the tracking environment (E) (see Fig. 1 (a)).

Environment Details: Adopting the RL framework from [21], we train an RL agent within the 3D diffusion MRI voxel space. The training process starts from seed voxels chosen within a 3D tract-specific mask (M), obtained from MRM. At any given voxel, the environment presents state (s_t) to the agent and rewards the agent's actions based on their alignment with the fODF peak, aiding in the learning of the optimized policy $\pi_{\theta opt}$. The tracking continues until the streamline exits the mask (M), surpasses a maximum length (l), or deviates significantly (>60°) from the previous tracking direction.

The **state** (s_t) is defined by 45 spherical harmonic (SH) coefficients and tracking mask (M) values from the current and six neighboring voxels, along with the four previous tracking directions, amounting to 334 dimensions ($7\times(45+1)+3\times4$). The predicted **action** (a_t) is a 3D vector representing tracking/fiber

direction. The action space of the environment is continuous, allowing the agent to explore a wide range of potential fiber directions, with values in the range [-1, 1]. The **reward** (r_t) at time-step t is given by the absolute dot product between the agent's predicted action (a_t) and the closest fODF peak (p_i), weighted by the dot product of the action (a_t) with the agent's previous tracking step ($\boldsymbol{u}_{t-1}$) (defined below).

$$r_t = \left| \max_{\boldsymbol{p}_i} (\boldsymbol{p}_i \cdot \boldsymbol{a}_t) \right| \times (\boldsymbol{a}_t \cdot \boldsymbol{u}_{t-1}) \tag{1}$$

Training Details: During the agent's exploration phase, the transitions (s, a, r, s') are recorded in a replay buffer for batch-wise policy optimization. We utilize the TD3 algorithm to train 7 tract-specific agents. It has an Actor and two Critic networks (along with their time delayed target networks).

The actor and critic networks are both fully-connected neural networks with two ReLU activated hidden layers of 1024 neurons each. The actor has a 334 dimensional input layer and 3-neuron tanh activated output layer, while the critic has a 337 dimensional input layer and a single neuron tanh output layer (similar to [21]). Each tract-specific RL agent is trained on five subjects from the TractoInferno dataset (1030, 1079, 1119, 1180, and 1198), for 50 batches (4096 episodes each) per subject, hence a total of 1,024,000 (250*4096) episodes. We train the TD3 agent in 5 different instances of the environment (E) specified by each subject's distinct diffusion data, fODF peaks, and tracking mask. Training is conducted at 7 seeds per voxel and a step-size of 0.375 mm, with fiber lengths between 20 mm and 200 mm. Maximum possible episode length is set to 530 (200/0.375). Other hyper-parameters include: learning rate: 8.56e-06, Discount factor (γ): 0.776, and Exploration noise (σ_{train}): 0.334.

3.3 T-RLF: Policy Refinement

This subsection involves the training steps of our T-RLF model. We train a GPT-based network, to learn a refined, $level-2$ policy ($\pi_{\phi opt}$) for tract-specific fiber generation. It is trained on the policy rollouts of the $level-1$ TD3 policy ($\pi_{\theta opt}$) (ref: Sect. 3.2) to interpret and generate fiber data within the agent's experience space. This is accomplished through a two-stage process: (a) Initially, the Tract-RLFormer undergoes a generic, tract-agnostic **pre-training.** (b) This is followed by **fine-tuning** for the downstream task of tract-specific generation. Together, these constitute the next three steps (out of five), namely training data generation (Fig. 1(b)) and the two-stage training process (Fig. 1(c, d)) of T-RLF. Each component of the training framework is discussed in detail below.

Unlike prior methods that generate fiber points by training on diffusion information along ground truth fiber streamlines, our network, T-RLF, learns from the sequence of state-action-reward (s, a, r) tuples (policy roll-outs) of a trained RL agent (Fig. 2). T-RLF is trained on trajectories derived from the policy roll-outs of seven tract-specific TD3 agents. Each trajectory is represented as $\boldsymbol{\tau = (R_0, s_0, a_0, R_1, ..., R_T, s_T, a_T)}$, where $\boldsymbol{R_t}$ is the scalar sum of rewards

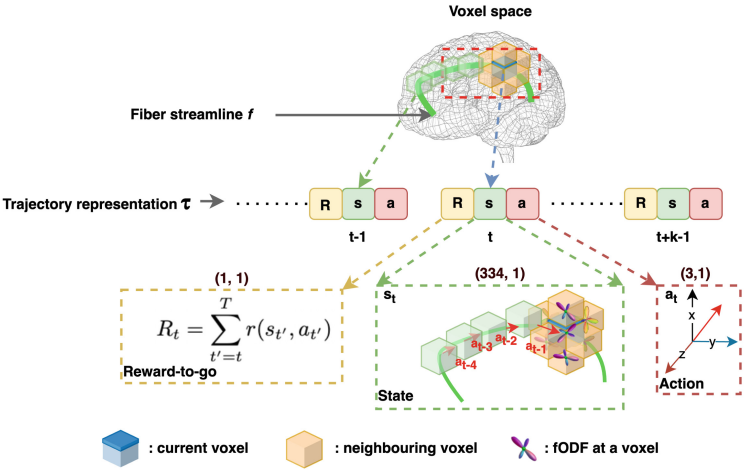


Fig. 2. Data Representation for T-RLF: Tract specific policy refinement using a trajectory-based approach in an RL agent's experience space. The figure illustrates a k length fiber streamline f in human brain voxel space, represented as a trajectory $\tau = (R_0, s_0, a_0, R_1, s_1, a_1, ..., R_k, s_k, a_k)$. Each point in the streamline corresponds to a state, action, and return-to-go tuple at a time-step t.

from time-step t to the episode's end, $\boldsymbol{s_t}$ is a 334-dimensional state vector, and $\boldsymbol{a_t}$ is a 3-dimensional action (see Sect. 3.2).

Training Data Generation: To generate training data trajectories, we initiate tracking for the 7 trained TD3 agents (see Sect. 3.2) on 5 training subjects from the TractoInferno dataset. For each of the 7 tracts, we save all tracking episodes (until termination) as (R, s, a) sequences, called trajectories. Tracking is conducted for all 5 subjects within their tract-specific masks using 7 seeds per voxel, resulting in a total of $\sum_{s=1}^{5} 7 \times n_{v_{s,i}}$ tract-specific trajectories for the i^{th} tract, where $n_{v_{s,i}}$ is the number of voxels in the i^{th} tract's mask for the s^{th} subject. From these, 50,000 trajectories are selected per tract, with 10,000 from each subject. Half of these (5,000) are the longest trajectories for that subject's i^{th} tract, while the other half represent the streamline variability of the tract. This yields a tract-specific dataset τ_i for each tract i used for model fine-tuning for downstream tasks. From the 350,000 ($7 \times 50{,}000$) trajectories of the seven tract-specific datasets (τ_i for $i = 1$ *to* 7), a total of 150,000 trajectories are selected. Half (75,000) of these are the longest trajectories, and the other half are randomly selected, resulting in a mixed tract dataset τ_{mix} used for generic tract-agnostic pre-training.

It should be noted that $n_{v_{s,i}}$ varies with the tract and subject. $n_{v_{s,i}}$ is the total number of voxels within the tracking mask for tract i (M_i) when aligned to the space of subject s. Moreover, the minimum and maximum length of trajectories in the τ_{mix} dataset (representative of all tracts) are 48 and 292 respectively. It is later used to determine the training parameter of GPT model.

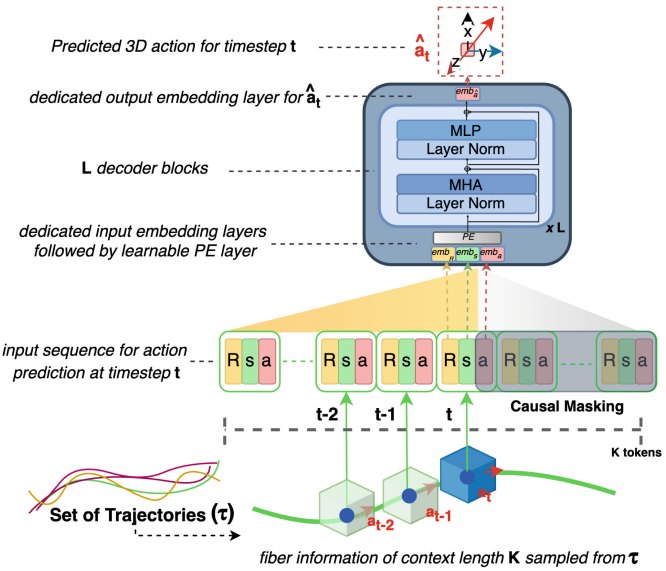


Fig. 3. Data Driven Policy Learning: Visual representation of training Tract-RLFormer for action prediction at time-step t, using context information from K length fiber (Sect. 3.3). The input sequence tuples $< R, s, a >$ are causally masked from a_t onwards and processed through embedding layers emb_R, emb_s, and emb_a, with a learnable positional encoding layer (PE). Embeddings are processed by L decoder blocks ($L = 3$ for pre-training, $L = 4$ for fine-tuning), incorporating Multi-Head Attention (MHA) and Multi-Layer Perceptron (MLP), to generate predicted action $\hat{a}_t$.

Model Architecture: Tract-RLFormer adopts the GPT architecture (as shown in Fig. 3) to model trajectories autoregressively [5]. The network consists of 4 decoder layers with 1 attention head each (n_heads=1), a context length (K) of 40, an embedding dimension (d) of 128, ReLU activation functions, and a dropout rate of 0.1. These parameters were selected after thorough experimentation presented in Sect. 4.3. It begins with a dedicated embedding layer of 128 dimensions (as shown in Fig. 3) for each component of the trajectory: state (s), action (a), and return-to-go (R). Subsequently, a trainable positional encoding layer processes the timestep sequence (of max_ep_len) as input, generating positional/timestep embeddings of dimensionality $d = 128$, where each timestep (t) has 3 tokens $< R_t, s_t, a_t >$. The maximum possible episode length (max_ep_len) controls length of episode. It is set to 530 because the maximum length of a fiber is 200 mm, equivalent to 530 steps for a TractoInferno subject (as 1 step corresponds to 0.375 mm; refer 3.4). If an episode exceeds 530 timesteps, it is truncated to this length. Embeddings for each component of the trajectory (state, action, return-to-go) are then combined and fed into the decoder layers. We utilize four decoder blocks, where each block includes a multi-headed self-attention mechanism followed by position-wise feed-forward networks. After processing

through the decoder blocks, the output is passed through an output embedding layer, from which we obtain the predicted action of dimension (3, 1).

Training Details: The proposed T-RLF model is trained to generate an optimal level-2 policy, ($\pi_{\phi opt}$), specifically tailored for tract-specific generation. It undergoes a two-stage training process, starting with general pre-training on mixed tract dataset (τ_{mix}), followed by tract-specific fine-tuning on tract-specific dataset (τ_i). The first three decoder layers are pre-trained over 0.15 million mixed trajectories (taken from τ_{mix}), containing a total of 30 million transitions for 30 iterations. Later the 4^{th} decoder layer is fine-tuned on the tract-specific trajectories buffer for 10 additional iterations. In each iteration, the model undergoes 10,000 training steps, each processing a $batch\_size$= 128 number of K-length trajectories. A batch of 128 tokens of $< R_t, s, a >$ are sampled from training data (τ) and stacked for a context length ($K = 40$) and fed as an input to the T-RLF. It passes through an embedding layer with 128 dimensions, and positional encoding is added, resulting in a ($128 \times 120 \times 128$) matrix and is processed by the 4 decoder layers with causal masking (Fig. 3). The decoder output is mapped through an output embedding layer to predict the action. Unlike the TD3 agent, T-RLF does not interact with the environment during its training process. Instead, it is trained entirely in an offline mode using only trajectory datasets (τ's). For the context length K, a 5-step loss (accounting for current and 2 steps in both forward and backward directions) is computed, aggregating the angular difference between predicted and actual action at each time-step.

$$L = \sum_{t=2}^{K-2} \left(\sum_{i=-2}^{2} \cos^{-1} \left(\mathbf{a}_{t+i} \cdot \hat{\mathbf{a}}_{t+i} \right) \right) \tag{2}$$

The learning of weights for $\pi_{\phi opt}$ is facilitated by this 5-step loss function, in order to generate more effective and robust actions. Here, $\boldsymbol{a}_{t+i}$ and $\hat{\boldsymbol{a}}_{t+i}$ are the true and predicted actions at $(t+i)^{th}$ timestep respectively.

Similar to [5], T-RLF training is conditioned to generate action (a_t) using return (R_t) at each timestep. During inference, R_t is initialized to an expert return value or the longest trajectory return. In our case, the longest trajectory length is 292, and since the maximum possible reward at each timestep is 1, we initialize R_t to 300 ($\sim$1x expert return). This was experimentally verified among various values: 100, 200, 300, 500, and 600. For model training, we employed the AdamW optimizer, set with a learning rate of 1e-4 and a weight decay of 1e-4.

3.4 T-RLF: Inference

The final step in our fiber tract generation method involves using the trained T-RLF models to perform tracking, followed by cleaning the resulting tracts. Having learnt the refined policy ($\pi_{\phi opt}$), T-RLF can function autonomously as a generic substitute for TD3 agent. **Consequently, it can independently perform fiber streamline generation in the same environment** (E) as detailed in Sect. 3.2, without relying on the original TD3 agent. Fiber generation (tracking) is executed within tract-specific masks obtained from MRM and

is initialised with 7 seeds per voxel, and R_t is set to $R_0 = 300$. Tracking step size is (empirically selected) and is dataset-specific, 0.375 mm for the TractoInferno, 0.468 mm for HCP, and 0.75 mm for the ISMRM dataset. At each step, the return-to-go (R_t) is reduced by the achieved reward and predicted action(a_t), new state (s'_t), and R_t are appended to the context window to serve as input for the next prediction. This auto-regressive process by Tract-RLFormer generates the fiber tract of interest. Finally, the tracts undergo a **Cleaning procedure** using a fast streamline search (FSS) [19] to eliminate any extraneous fibers, by comparing the predicted tract with the atlas reference tracts (representing general anatomical structure). Our tracts are confined to masks generated by the MRM module, tailored to each subject's fiber orientation. This approach ensures that tract generation remains confined to regions proximate to the actual neural fibers of the subject, thus mitigating the risk of false positives. Consequently, we can perform a high radius search using FSS, without incurring a major risk of high overreach. This high radius search ensures that accurate fibers are not discarded based on minor deviations from atlas tracts.

Performance Parameters: In order to evaluate the quality of our generation, the Ground Truth tract is aligned to Montreal Neurological Institute (MNI) space using Advanced Normalization Tools (ANTs) [1], facilitating comparison with our cleaned tracts that are already in MNI space. The Dice (D), Overlap (OvL), and Overreach (OvR) scores (similar to [20,21]) are then computed against the ground truth tract and are reported in Sect. 4. The Dice score assesses both the accurate coverage and the minimization of extraneous extensions beyond the ground truth area, where values near 1 signify a high similarity level. Overlap measures the intersection of the generated tract with the ground truth, while Overreach indicates how much the generated tract exceeds the ground truth, with lower scores suggesting greater precision.

4 Results and Discussion

In this section, we present the outcomes of our evaluation of tract-specific T-RLF models under various experimental setups, including comparative analysis, generalization performance, and an ablation study. We trained TD3 and T-RLF models, on eight tracts- seven principal white matter tracts (refer Sect. 3) and OR tract (for analysis in 4.1) using five **train subjects** (id: 1030, 1079, 1119, 1180, and 1198) of the TractoInferno dataset and reported their performance on various test subjects across different datasets in subsequent subsections. Additionally, we assess their effectiveness relative to supervised approaches and traditional tractography methods that do not incorporate learning.

4.1 Comparative Analysis

This section provides a comparative analysis of our model, T-RLF, against supervised learning, traditional tractography, and state-of-the-art (SOTA) reinforcement learning (RL) algorithms, using Dice scores to evaluate performance across

Table 2. Comparison of mean Dice scores for the OR, PYT, and CC tracts for subject 1006 from TractoInferno dataset. Supervised learning scores are from [16]; RL-based scores, with std. dev., are from [20]. The last 2 rows includes scores for T-RLF and TD3, evaluated using our tract-specific approach. The highest and second highest scores are highlighted in green and red, respectively. '*' denotes tract-specific setting for methods.

Algorithm	OR	PYT	CC
DET-SE	0.569	0.665	0.658
DET-Cosine	0.598	0.708	0.646
Prob-Sphere	0.599	0.695	0.648
Prob-Gaussian	0.542	0.723	0.668
Prob-Mixture	0.436	0.522	0.614
DET	0.516	0.475	0.345
PROB	0.549	0.740	0.590
PFT	0.644 ± 0.136	0.753 ± 0.010	0.827 ± 0.008
VPG	0.369 ± 0.135	0.434 ± 0.128	0.428 ± 0.182
A2C	0.225 ± 0.108	0.323 ± 0.082	0.222 ± 0.025
ACKTR	0.397 ± 0.171	0.559 ± 0.028	0.584 ± 0.054
TRPO	0.330 ± 0.154	0.498 ± 0.062	0.594 ± 0.048
PPO	0.440 ± 0.187	0.619 ± 0.042	0.650 ± 0.028
DDPG	0.612 ± 0.063	0.630 ± 0.045	0.731 ± 0.006
TD3	0.555 ± 0.097	0.603 ± 0.045	0.688 ± 0.035
SAC	0.598 ± 0.098	0.658 ± 0.028	0.753 ± 0.010
SAC Auto	0.608 ± 0.088	0.655 ± 0.032	0.747 ± 0.019
DET*	0.648	0.752	0.713
PROB*	0.652	0.765	0.731
TD3*	0.644	0.764	0.720
T-RLF (Ours)	0.673	0.772	0.738

three major white matter bundles: PYT, OR, and CC. As presented in Table 2, all methods are tested on **subject 1006** from the TractoInferno dataset (similar to [20,21] for fair comparison). For the first and third tabular subparts of Table 2, the models are trained on ISMRM data. The second subpart does not involve training (classical methods). These 3 subparts are assessed using whole-brain tractography and segmentation [16] [20]. Additionally, the last subpart details the performance of our T-RLF and the TD3 model, where T-RLF was specifically trained on trajectories derived from the TD3 agent.

In Table 2, our framework outperforms the state-of-the-art method (PFT) for PYT and OR tracts, demonstrating its robustness in tract-specific tractography. Additionally, T-RLF shows comparable performance to state-of-the-art RL algorithms for CC tract.

Furthermore, the TD3 agent demonstrates markedly improved performance within our tract-specific generation framework. Testing of tract-specific TD3 on the ISMRM or HCP datasets cannot be conducted due to the absence of the evaluated tracts in these datasets. However, the enhancement in TD3's performance in our tract-specific setting can be attributed to the training approach rather than dataset consistency. This is evidenced by TD3's comparable or superior performance on different tracts across the ISMRM and HCP datasets, as detailed further in Tables 3, 4.

Moreover, dice scores for DET and PROB improved for all tracts in the tract-specific setting, especially for CC, where DET increased by 106.67% (0.345 to 0.713) and PROB by 23.89% (0.590 to 0.731). The enhanced tracking performance of DET and PROB, despite not being trained, is indicative of the effectiveness of our tract-specific masks. Also, in the whole-brain setting, there is a huge difference between DET (0.475) and PROB (0.740) scores on the PYT tract (Table 2), whereas this gap is significantly smaller in the tract-specific setting (marked with '*'), where the tract-specific performance of DET* (0.752) and PROB* (0.765) align closely with each other and with the T-RLF and TD3 methods. The consistency and stability observed for these classical methods are attributed to our tract-specific approach.

4.2 Generalization Performance Evaluation

In this section, we present the performance evaluation of our T-RLF model across three distinct datasets (Tables 3, 4), demonstrating its effectiveness and generalizability. The averaged results include analyses across five **test subjects** in the TractoInferno (TtoI) dataset (id: 1160, 1078, 1159, 1061, and 1171), four from the HCP dataset (id: 930449, 992774, 959574, and 987983), and one from the ISMRM dataset. A visual comparison across datasets and subjects is presented in Fig. 4(a). We also compare the performance of T-RLF with classical algorithms, which were employed using tract-specific masks, and the tract-specific TD3 agent, from which the training data for T-RLF was derived (Fig. 4(b)).

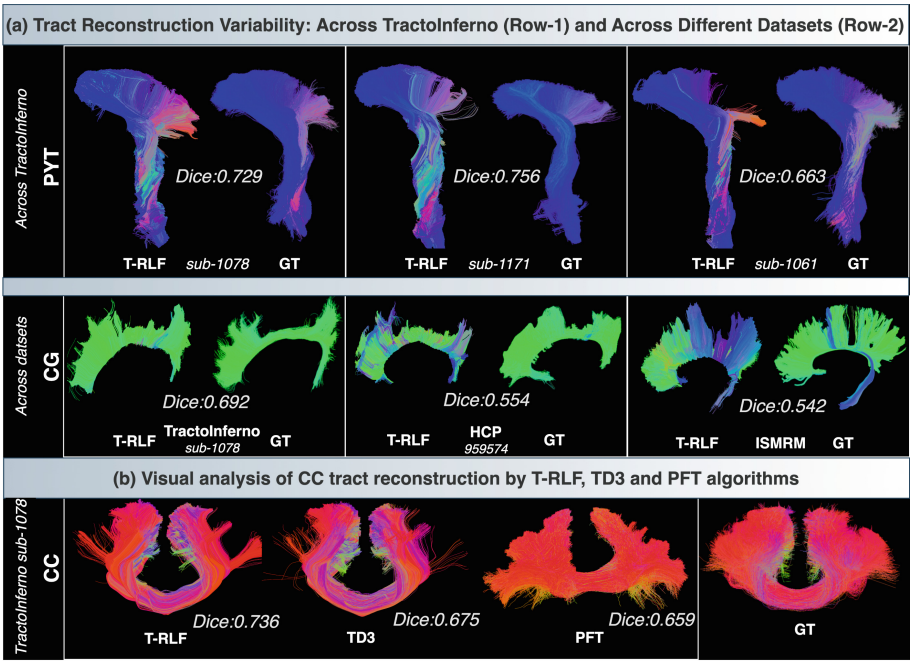


Fig. 4. Visual comparison of reconstructed tracts illustrating (a): Intra-dataset variability, Inter-dataset variability, and (b): Variability across tracts reconstructed by different algorithms. The depicted tracts include the left PYT, CG, and a part of CC. The algorithms evaluated in bottom section of figure are T-RLF (ours), TD3, and PFT.

In Table 3, we see that T-RLF model displays a notable generalization performance. Interestingly, the classical deterministic (DET) and probabilistic (PROB) methods exhibit slightly better performance than learnable methods in some cases (Tables 3,4).

As previously mentioned in Sect. 4.1, the consistency observed in Tables 3, 4 for the classical methods (DET and PROB) is due to our tract-specific approach. This improvement and stabilization may be attributed to the elimination of premature termination issues in narrow and deep WM regions, as described in [10], facilitated by the refined spatial exploration enabled by MRM in our tract-specific approach. It can be observed from Tables 2 and 4, that the performance of PFT declined in the tract-aware setting, dropping from 75% to 66.2% in the PYT and from 82% to 55% in the CC (refer Table 4). This decline can be attributed to use of Continuous Map Criterion (CMC) as a stopping criterion for fiber tracking. The CMC terminates fiber tracking based on Partial Volume Estimate (PVE) maps, allowing tractography to continue until the streamline correctly stops in the gray matter. This approach may generate fibers beyond our tract-specific masks, leading to increased overreach (see Fig. 4(b))

Table 3. Performance metrics(in %) for the CG and AF tracts, trained on the TractoInferno dataset and tested across multiple datasets to evaluate generalization. Tracking is performed using our proposed tract-specific generation method. A dash ('-') indicates the absence of ground-truth tracts in the corresponding dataset, precluding evaluation.

Dataset	Algo.	Cingulum (CG)						Arcuate Fasciculus (AF)					
		Left			Right			Left			Right		
		Dice	OvL	OvR	Dice	OvL	OvR	Dice	OvL	OvR	Dice	OvL	OvR
HCP	T-RLF	53.3	42.5	16.6	45.6	33.7	13.9	61.8	51.2	13.4	41.8	27.9	5.40
	TD3	53.0	42.3	16.9	45.2	33.6	14.3	61.6	51.0	13.7	41.6	27.7	5.60
	DET	55.2	46.4	21.5	52.6	41.7	16.3	62.8	52.5	14.0	43.9	30.0	6.6
	PROB	57.6	51.8	27.9	56.1	45.5	16.6	65.5	57.9	18.3	47.4	33.7	8.6
	PFT	67.3	55.2	7.8	59.6	45.4	6.4	71.3	71.9	29.7	69.9	71.6	33.3
TtoI	T-RLF	61.0	56.8	28.6	56.5	52.6	34.9	52.7	45.1	27.8	39.5	36.5	49.8
	TD3	60.0	55.1	27.3	54.9	49.8	32.4	51.8	44.3	28.2	38.4	34.9	46.9
	DET	61.2	58.8	32.3	58.2	54.7	33.7	54.6	46.3	24.7	45.4	41.7	46.9
	PROB	67.9	69.1	33.6	64.7	64.6	36.7	62.3	57.2	27.2	50.3	48.3	50.9
	PFT	55.9	48.8	25.4	54.5	51.3	38.6	62.8	60.1	31.5	53.9	62.8	88.0
ISMRM	T-RLF	54.2	46.6	25.5	52.8	44.1	23.1	-	-	-	-	-	-
	TD3	53.1	44.8	23.7	51.2	41.6	21.1	-	-	-	-	-	-
	DET	57.5	51.9	28.5	57.7	52.4	29.2	-	-	-	-	-	-
	PROB	61.1	59.3	35.1	64.0	65.4	39.0	-	-	-	-	-	-
	PFT	55.4	49.4	28.9	57.3	49.4	22.9	-	-	-	-	-	-

and consequently lower Dice scores. Furthermore, fibers generated outside the tracking mask may be erroneous and subsequently filtered or cleaned via FSS, resulting in a lower OvL score.

Table 4. Results are presented for the left and right parts of PYT and a segment of CC on the TractoInferno (TtoI) dataset. Tracking for all algorithms is conducted using our proposed tract-specific generation method.

Dataset	Algo.	Pyramidal Tract (PYT)						Corpus Callosum (CC)		
		Left			Right					
		Dice	OvL	OvR	Dice	OvL	OvR	Dice	OvL	OvR
TtoI	T-RLF	70.3	64.1	17.2	70.1	63.1	16.9	70.4	71.2	32.6
	TD3	69.4	62.5	15.9	69.2	61.4	15.8	68.1	64.8	26.1
	DET	72.7	79.3	38.8	70.3	76.2	40.7	70.1	72.6	35.8
	PROB	77.6	79.5	25.3	74.8	72.5	21.3	72.6	76.4	36.7
	PFT	66.2	55.7	12.4	65.9	57.8	17.4	54.9	51.2	36.1

Summarization: In summary, our results demonstrate that we surpass supervised methods (Table 2). Additionally, we consistently outperform the TD3 model (Tables 2, 3 and 4), which served as the basis for training T-RLF. Notably, our tract-specific setting not only improves TD3 performance but also the performance of classical methods like DET and PROB compared to the whole-brain setting. This suggests a promising new direction of data driven policy learning for tract specific fiber generation in limited ground truth scenarios that can naturally scale up effectively.

4.3 Ablation Study

We conducted an ablation study to determine the optimal configuration for our T-RLF model. The evaluation presented in Table 5, identified the best architecture with n_heads=1, K=40, and an embedding dimension of d=128. This study highlights the importance of a larger context in tractography, illustrating how a broader temporal receptive field can enhance the model's ability to generate accurate fiber tracts.

Table 5. Dice scores (in %) averaged over 7 tracts of subject 1006 from TractoInferno dataset, at different values of T-RLF parameters: number of attention heads (n_heads), context length (K), and embedding dimension (d). Best score is in **bold**.

	$K = 20$		$K = 30$		$K = 40$	
	$d = 128$	$d = 512$	$d = 128$	$d = 512$	$d = 128$	$d = 512$
$n\_heads = 1$	64.7	65.2	66.2	67.3	**68.6**	67.6
$n\_heads = 2$	63.4	66.3	66.2	67.5	68.0	68.2

We also examined the impact of two key components: Mask Refinement Module (MRM) discussed in Sect. 3.1, and the tract-specific policy fine-tuning as detailed in Sect. 3.3. Table 6 reports the results for the T-RLF network trained over TractoInferno dataset. We have observed that initial tracking masks led to a significant overreach (OvR), extending beyond actual region of interest. This OvR was notably reduced after incorporating MRM, leading to improved Dice and overlap metrics across all tracts. Furthermore, fine-tuning the network specific to each tract allowed it to learn better and robust tract-specific diffusion characteristics, resulting in additional improvements in the performance metrics.

Table 6. Average performance metrics (in %) obtained using Tract-RLFormer highlight the impact of the MRM on test subjects from the TractoInferno dataset. The table also compares the performance of the pre-trained network with the fine-tuned network post MRM application, illustrating the effect of policy fine-tuning on the same dataset.

Tract	Without MRM			With MRM					
	Dice	OvL	OvR	Pre-trained			Fine-tuned		
				Dice	OvL	OvR	Dice	OvL	OvR
PYT	44.1	30.4	5.6	65.9	55.1	11.8	70.3	67.2	24.7
CG	41.1	45.4	80.3	51.5	45.7	30.3	58.7	54.7	31.7
AF	34.2	34.4	65.1	45.8	39.5	36.1	46.1	40.8	38.8
CC	58.9	59.0	41.9	66.2	59.9	20.8	70.4	71.2	32.6

5 Conclusion

Tractography can be an essential tool in neuroimaging, enabling the detailed mapping of neural pathways crucial for both clinical and research applications. Our work significantly advances this field by introducing a data driven Tract-RLFormer framework which is a tract-specific, transformer-based network integrating supervised and reinforcement learning paradigms. A distinctive feature of our Tract-RLFormer is its ability to train within the reinforcement learning experience space, independent of ground truth fibers. The fine-tuning stage of our model focuses and refines its capabilities in generating the tracts of interest. This approach demonstrates its excellent generalization performance across various datasets as well as scalability. Our data-driven approach has the potential to utilize data from any reinforcement learning agents trained in diverse neuroimaging environments. Moreover, our innovative tract-specific modeling approach simplifies the reconstruction process by directly generating the target tract, thus avoiding the complex and error-prone segmentation step.

Acknowledgment. This research was supported by SERB Core Research Grant Project No: CRG/ 2020/005492, IIT Mandi.

References

1. Avants, B.B., et al.: Advanced normalization tools. Insight J **2**(365), 1–35 (2009)
2. Basser, P.J.: Fiber-tractography via diffusion tensor MRI (DT-MRI). In: Proceedings of the 6th Annual Meeting ISMRM, Sydney, Australia, vol. 1226, p. 14 (1998)
3. Benou, I., Riklin Raviv, T.: Deeptract: A probabilistic deep learning framework for white matter fiber tractography. In: MICCAI: Shenzhen, China, October 13–17, 2019, pp. 626–635. Springer (2019)
4. Berman, J.I., Chung, S., Mukherjee, P., Hess, C.P., Han, E.T., Henry, R.G.: Probabilistic streamline q-ball tractography using the residual bootstrap. Neuroimage **39**(1), 215–222 (2008)
5. Chen, L., et al.: Decision transformer: reinforcement learning via sequence modeling. NeurIPS **34**, 15084–15097 (2021)

6. Dosovitskiy, A., et al.: An image is worth 16x16 words: transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
7. Essayed, W.I., Zhang, F., Unadkat, P., Cosgrove, G.R., Golby, A.J., O'Donnell, L.J.: White matter tractography for neurosurgical planning: A topography-based review of the current state of the art. NeuroImage: Clinical **15**, 659–672 (2017)
8. Fillard, P., Poupon, C., Mangin, J.F.: A novel global tractography algorithm based on an adaptive spin glass model. In: MICCAI, pp. 927–934. Springer (2009)
9. Fujimoto, S., Hoof, H., Meger, D.: Addressing function approximation error in actor-critic methods. In: ICML, pp. 1587–1596. PMLR (2018)
10. Girard, G., et al.: Towards quantitative connectivity analysis: reducing tractography biases. Neuroimage **98**, 266–278 (2014)
11. Jumper, J., Evans, R., Pritzel, A., Green, T., et al.: Highly accurate protein structure prediction with alphafold. Nature **596**(7873), 583–589 (2021)
12. Maier-Hein, K.H., et al.: The challenge of mapping the human connectome based on diffusion tractography. Nat. Commun. **8**(1), 1349 (2017)
13. Neher, P.F., Côté, M.A., Houde, J.C., Descoteaux, M., Maier-Hein, K.H.: Fiber tractography using machine learning. Neuroimage **158**, 417–429 (2017)
14. Neher, P.F., et al.: A machine learning based approach to fiber tractography using classifier voting. In: MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part I 18, pp. 45–52. Springer (2015)
15. Poulin, P., et al.: Learn to track: deep learning for tractography. In: MICCAI 2017: 20th International Conference, Quebec City, QC, Canada, September 11-13, 2017, Proceedings, Part I 20, pp. 540–547. Springer (2017)
16. Poulin, P., et al.: Tractoinferno-a large-scale, open-source, multi-site database for machine learning DMRI tractography. Sci. Data **9**(1), 725 (2022)
17. Rheault, F.: Population average atlas for recobundlesx (2023). https://doi.org/10.5281/zenodo.7950602
18. Rheault, F., et al.: Bundle-specific tractography with incorporated anatomical and orientational priors. Neuroimage **186**, 382–398 (2019)
19. St-Onge, E., Garyfallidis, E., Collins, D.L.: Fast streamline search: an exact technique for diffusion MRI tractography. Neuroinformatics **20**(4), 1093–1104 (2022)
20. Théberge, A., Desrosiers, C., Boré, A., Descoteaux, M., Jodoin, P.M.: What matters in reinforcement learning for tractography. MIA **93**, 103085 (2024)
21. Théberge, A., et al.: Track-to-learn: a general framework for tractography with deep reinforcement learning. MIA **72**, 102093 (2021)
22. Van Essen, D.C., et al.: The human connectome project: a data acquisition perspective. Neuroimage **62**(4), 2222–2231 (2012)
23. Vaswani, A., et al.: Attention is all you need. In: Advances in Neural Information Processing Systems, vol. 30 (2017)
24. Wegmayr, V., Buhmann, J.M.: Entrack: probabilistic spherical regression with entropy regularization for fiber tractography. Int. J. Comput. Vis. **129**(3), 656–680 (2021)
25. Wegmayr, V., Giuliari, G., Holdener, S., Buhmann, J.: Data-driven fiber tractography with neural networks. In: 2018 IEEE 15th international symposium on biomedical imaging (ISBI 2018), pp. 1030–1033. IEEE (2018)
26. Yeh, F.C., et al.: Population-averaged atlas of the macroscale human structural connectome and its network topology. Neuroimage **178**, 57–68 (2018)
27. Zhou, H., Zhang, et al.: Informer: beyond efficient transformer for long sequence time-series forecasting. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 35, pp. 11106–11115 (2021)

Detecting Concept Shifts Under Different Levels of Self-awareness on Emotion Labeling

HyoSeon Choi[1], Dahoon Choi[2], Netiwit Kaongoen[4], and Byung Hyung Kim[1,3(✉)]

[1] Department of Electrical and Computer Engineering, Incheon, Republic of Korea
[2] Department of Computer Engineering, Incheon, Republic of Korea
[3] Department of Artificial Intelligence, Inha University, Incheon, Republic of Korea
bhyung@inha.ac.kr
[4] School of Computing, KAIST, Daejeon, Republic of Korea

Abstract. Generalizing deep learning for all requires individual self-assessment. However, the quality of ground-truth labels depends on the annotators' self-awareness. Real-world datasets inevitably experience the *Concept Shift* problem. Recent advances in Out-of-distribution (OOD) detection have received much attention due to its ability to alleviate distribution shift problems by distinguishing between anomalous and in-distribution(ID) data samples. Existing approaches underlie pre-trained ID models learned with class-balanced data. However, this assumption makes the methods incapable when the ID models are trained with inter- and intra-class variance depending on user characteristics, such as gender, culture, and genetics. We present an OOD detection framework. Our system builds a generalized ID model by extracting high-quality data from high-dimensional neural activities considering individuals' cognitive and perceptional ability to evaluate self-assessments. The proposed system detects and removes abnormal pairs of data and labels to enhance model performance by considering the maximum softmax probability approach. Experimental results on public EEG datasets in emotion recognition demonstrate the superiority of our method despite the non-stationary nature of EEG signals. The codes are available at https://github.com/affctivai/coglier.

Keywords: Concept Shift · EEG · Emotion · Labeling · Self-awareness

1 Introduction

The development of reliable deep learning-based machine systems has garnered increasing attention. This development necessitates the ability to generalize

This work was supported in part by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (RS-2023-00229074, RS-2022-00155915), in part by the National Research Foundation of Korea (NRF) grant funded by the Korea government (No. 2021R1C1C2012437), and in part by INHA UNIVERSITY Research Grant.

A. Antonacopoulos et al. (Eds.): ICPR 2024, LNCS 15313, pp. 276–291, 2025.
https://doi.org/10.1007/978-3-031-78201-5_18

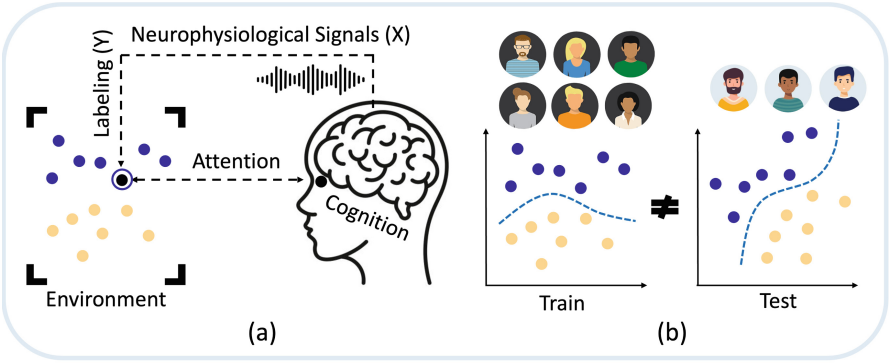


Fig. 1. *Concept Shift* under Different Levels of Self-awareness on Emotion Labeling. (a) Neurophysiological signals (X) objectively reflect human cognitive states, representing individual levels of labeling (Y) awareness. (b) Human-annotated datasets inevitably experience *Concept Shift* ($P_{\text{train}}(Y|X) \neq P_{\text{test}}(Y|X)$) due to the inherent ability in individual cognition and perception to understand and categorize their environment.

their predictive capabilities to new data, which often includes both known and unknown classes, in real-world scenario [11,15]. While deep learning models have increased such capability with high performance in various applications, they tend to be overconfident in the unknown classes [18]. These overconfident behaviors by confusing known and unknown classes affect the model generalization adversely in real-world environments.

In light of the recent success in out-of-distribution (OOD) detection with deep neural networks, several inference-time or post-hoc methods have been proposed to identify unknown classes while correctly classifying known ones [25]. Current methods learn in-distribution (ID) data to recognize OOD data, adopting the open-world assumption, in which test samples can be OOD drawn from either different classes or a different domain [25]. The assumption can be formulated as *Distribution Shift*, which occurs when the joint distribution at training is not equal to that at testing; that is, $P_{\text{train}}(X, Y) \neq P_{\text{test}}(X, Y)$. Existing approaches underlie pre-trained ID models trained with class-balanced data for the model generalization. However, this assumption makes the above methods incapable when their ID model is trained with inter- and intra-class variance [12]. Unfortunately, this incapability is natural in real-world applications that require individual self-assessment [4,17]. Carrying out precise self-assessment, such as data labeling, is almost impossible in such scenarios. Human annotation on extensive data is too costly. Moreover, the quality of ground-truth labels depends on the annotators' level of self-awareness [10,24]. Recent advances in wearable technologies enable the record of neurophysiological activity patterns, which vary depending on the characteristics of the users, such as gender, culture, and genetics, in the wild [20]. Neurophysiological signals objectively reflect human cognitive and emotional states in response to the human body's central nervous system (CNS) and autonomic nervous system (ANS). Among the signals, recent studies have increasingly utilized EEG as inputs (X) to investigate cognition-emotion

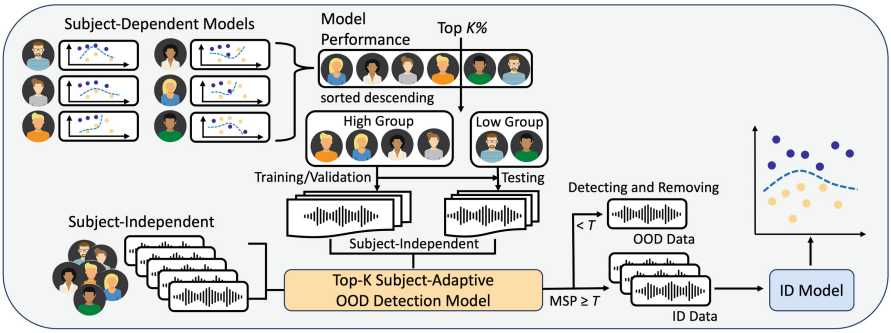


Fig. 2. Overview of the Proposed Framework. The removal process of data corresponding to Out-of-Distribution (OOD) is added. (1) A model is trained and evaluated for each subject, followed by an accuracy distribution analysis. Subjects are then categorized into High/Low Groups based on Top-K%. (2) High Group data are shuffled (subject-independent) to train the OOD Detection Model (ODM). The detection performance of ODM is evaluated with paired 'High Group's testing data' and 'Low Group data.' (3) Training data is filtered through the pre-trained ODM to obtain MSP. Samples with MSP above a threshold (T) are classified as In-Distribution (ID), and an ID Model is trained accordingly. Otherwise, the samples are considered as OOD data to be removed.

interactions by analyzing high-dimensional neural activities in the brain. [14]. Since data labeling (Y) is a primary cognitive process, the signals (X) represent individual levels of labeling awareness, which in turn have a decisive effect on the quality of labels (Fig. 1).

Hence, real-world datasets inevitably experience the *Concept Shift* problem, which occurs when $P_{\text{train}}(Y|X) \neq P_{\text{test}}(Y|X)$ [25]. Particularly, building emotion-aware systems suffers from inter- and intra-subject variability problems in detecting emotional changes due to the inherent ability in individual cognition and perception to understand and categorize their environment [23]. The main challenge is that one needs to detect individual levels of labeling awareness, which causes *Concept Shift*. Besides, whereas most OOD research is based on well-curated data in computer vision and natural language processing, other non-visual modalities have inherent difficulties in visualization and interpretation for user-curated ID data collection [9].

To solve the above problem, we propose an OOD detection framework in emotion recognition (Fig. 2). Our system builds a generalized ID model by extracting high-quality of subject-wise data from high-dimensional neural activities considering individuals' cognitive and perceptional ability to evaluate self-assessments. The proposed system detects and removes abnormal pairs of data and labels to alleviate the *Concept Shift* problem by considering the maximum softmax probability (MSP) approach from the output layer of a deep learning model as a confidence measure to refine the dataset [8].

We conduct extensive experiments to show the effectiveness of the proposed system. Experimental results on public EEG datasets in emotion recognition

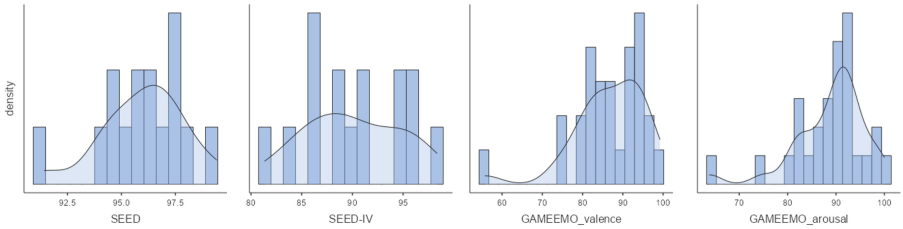


Fig. 3. Density plots showing the accuracy distributions for SEED, SEED-IV, and GAMEEMO datasets. The right-skewed distributions with long left tails show that most subjects had high accuracy, with a few having significantly lower accuracy.

demonstrate the superiority of our method for learning discriminative ID features of EEG signals despite their non-stationary nature. Furthermore, we investigate the difference between the ID and OOD identified by the proposed model based on topological distributions. This approach provides neuroscientific evidence that people with low cognitive levels correlate highly to provide abnormal data and label pairs. The analysis gives subject-independent insights, paving the way for enhanced emotion recognition systems by serving as a high-level outlier removal technique for supervised learning.

2 Methodology

Data labeling is a primary cognitive process that allows individuals to understand and categorize their environment. However, the labels assigned to the data can be inconsistent due to biased judgments or subjective interpretations. This inconsistency in data distribution across users leads to a 10–20% decrease in performance for subject-independent models compared to subject-dependent ones [14]. Motivated by the inconsistency, we make the following assumptions:

- *Concept Shift* occurs when irregular data and label distributions are present.
- People with low cognitive levels provide abnormal pairs of data and labels.
- Subject-dependent models would perform worse when those with lower cognitive levels conduct data labeling.

High- and Low-cognitive ability to evaluate self-assessments. Under the assumptions, we consolidate cross-subject data samples in ID from subject-dependent data, dividing them into two levels on the basis of self-awareness on labeling. While common studies have treated heterogeneous datasets as ID and OOD, we aim to separate the two distributions from subject-dependent datasets: one with High Label-Reliability (High Group) and another with Low Label-Reliability (Low Group).

Top-K Subject-Adaptive OOD Detection. The variability in accuracy between subjects is a characteristic feature of users' involved real-world applications. Besides, it can also be attributed to *Concept Shift* in the context of

supervised learning. Top-K% subject-dependent models sorted by their accuracies in descending order are selected. All data samples associated with the selected models are grouped in the High Group, otherwise grouped in the Low Group. Hence, the two groups are subject-independent and disjoint. We designate the Top 85% (Q_{85}), underlying a 10–20% decrease in performance for subject-independent systems compared to subject-dependent ones. Furthermore, our observation across various datasets and models suggested the value as a reasonable guideline (Fig. 3), although not fixed because of inter- and intra-subject variability. That is, the criterion is not strictly fixed at 85%; it can be adjusted by 1–2 subjects based on the distribution in Fig. 3, at the experimenter's discretion.

Detecting and Removing OOD Data. To generalize further, we attribute the degradation in model performance to broken correlations between X (data) and Y (label) and aim to apply the OOD detection methods described. Thus, our core premise is that the database contains non-corresponding (abnormal) data-label pairs, and identifying and eliminating these as OOD will likely enhance the classification performance.

Typically, OOD detection is done in an inference process using the confidence score of the model—as intuitively expected, higher confidence tends to be ID, and lower confidence tends to be OOD. The proposed method leverages this for detecting and removing OOD data samples using the MSP with the threshold T. If MSP $< T$, our framework detects the sample as an outlier, which should be removed. Our framework computes the MSP by OOD detection model trained from the subject-independent High Group. It should be noted that High Group data is used for training, validation, and evaluation while the Low Group data is used for evaluation only. The proposed framework is summarized in Algorithm 1.

3 EXPERIMENTS

3.1 Datasets and Data Preprocessing

In this study, we evaluated our framework for EEG-based emotion classification tasks with three public EEG datasets.

- SEED [28] contains EEG signals induced from 45 maximum 4-minute video clips, eliciting three emotions (positive, negative, and neutral). Fifteen subjects participated in 45 trials over three sessions, resulting in 675 data samples. In each trial, the subjects watched a single video, and EEG signals were collected using a 62-channel ESI NeuroScan System. The EEG signals were filtered between $1 \sim 75$ Hz, and the sampling frequency was 200 Hz. A sliding window method was applied with a window size of $400 (= 2\,\text{s} \times 200)$, resulting in a final dataset of 152,055 samples.
- SEED-IV [27] contains 64-channeled EEG signals recorded by the same equipment as SEED. Fifteen participants watched 72 maximum 2-minute video clips, reporting their emotions among four categories (happiness, sadness, fear, and neutral). They participated in a total of 72 trials over three sessions, resulting in a total of 1,080 data samples. EEG signals were filtered

between $1 \sim 75$ Hz, and the sampling frequency was 200 Hz. A sliding window of 400 ($= 2\,\text{s} \times 200$) was applied, resulting in a final dataset size of 150,765 samples.

– GAMEEMO [1] is an EEG dataset obtained by playing games of four genres (boring, calm, horror, funny) for 5 min each, totaling 20 min of EEG data. Twenty-eight participants evaluated their emotions, annotating discrete valence and arousal ratings on scales from 1 to 9 using the Self-Assessment Manikin (SAM) assessment tool [16]. Valence represents the degree of positivity or negativity, and arousal indicates the level of emotional activation,

Algorithm 1: Top-K Subject-Adaptive OOD Detection Framework

Input: Data for each subject i: $D_{\text{sub}_i} = \{(X, y)\}$, where $y \in \{1, 2, \ldots, C\}$
Output: $D_{\text{ID}}^{\text{train}}, D_{\text{OOD}}^{\text{train}}, \theta_{\text{ODM}}, \theta_{\text{ID}}$
Initialization: Split data D_{sub_i} into $D_{\text{sub}_i}^{\text{train}}$ and $D_{\text{sub}_i}^{\text{test}}$ for all $i = 1, 2, \ldots, N$

1. Subject-dependent Evaluation
for *each subject* $i = 1, 2, \ldots, N$ **do**
  Train and evaluate with parameter θ_i
Sort accuracies $\{\text{acc}_1, \text{acc}_2, \ldots, \text{acc}_N\}$, set k to the index of Q_{85}
Divide into: High Group $\{\text{acc}_1, \ldots, \text{acc}_{k-1}\}$ and Low Group $\{\text{acc}_k, \ldots, \text{acc}_N\}$

2. OOD Detection Model (ODM) Training
Train ODM using $D_{\text{High}}^{\text{train}} = \{D_{\text{sub}_1}^{\text{train}}, D_{\text{sub}_2}^{\text{train}}, \ldots, D_{\text{sub}_{k-1}}^{\text{train}}\}$ to obtain θ_{ODM}
Evaluate ODM using $D_{\text{Low}}^{\text{train}} = \{D_{\text{sub}_k}^{\text{train}}, D_{\text{sub}_{k+1}}^{\text{train}}, \ldots, D_{\text{sub}_N}^{\text{train}}\}$

3. Data Refinement
for *each* $(X, y) \in D^{train}$ **do**
  Calculate MSP:

$$\text{MSP} = \max\left(\frac{e^{z_{X,j}}}{\sum_{j=1}^{C} e^{z_{X,j}}}\right)$$

  if MSP $\geq$ *threshold* **then**
    Add (X, y) to $D_{\text{ID}}^{\text{train}}$
  else
    Add (X, y) to $D_{\text{OOD}}^{\text{train}}$
return $D_{\text{ID}}^{\text{train}}, D_{\text{OOD}}^{\text{train}}$

4. ID Model
Train the ID Model using $D_{\text{ID}}^{\text{train}}$ to obtain θ_{ID}
for *each* $(X, y) \in D^{test}$ **do**
  Calculate MSP using θ_{ODM}
  if MSP $\geq$ *threshold* **then**
    Evaluate (X, y) using θ_{ID}

with higher values indicating more positive emotions and higher arousal levels. EEG signals acquired through 14-channel EMOTIV EPOC+ were filtered from 0.16 to 43 Hz, with a sampling frequency of 128 Hz. By applying a sliding window of 256 (= 2 s × 128) with an overlap of 128, the final dataset size was 33,264 samples.

Each dataset was divided into training data (90%) and test data (10%) within each subject, with 10% of the training data used for validation. For ODM training, the High Group data was split 9:1, with the smaller portion used for ODM detection performance.

3.2 Baselines and Model Comparisons

Three deep learning models were employed for generality: CCNN [26] and TSception [7], both based on convolutional neural networks (CNNs), and the graph-based Dynamic Graph Convolutional Neural Network (DGCNN) [22]. We further compared the model performance in outlier detection with the Riemannian Potato (RP) method, a multivariate adaptive method for identifying artifacts in continuous data [3]. The principle of the RP is to represent clean signals by estimating a reference and a measure of dispersion (z-score) for each epoch.

3.3 Experimental Setup

EEG Feature Extraction We used two widely used EEG features: Differential Entropy (DE) and Power Spectral Density (PSD) [2]. DE divides EEG signals into four frequency bands (theta (4 $\sim$ 7 Hz), alpha (8 $\sim$ 13 Hz), beta (14 $\sim$ 31 Hz), and gamma (32 $\sim$ 49 Hz) and calculates the differential entropy for each band. Extracted DE features are transformed into a grid (9 × 9) array to match the electrode locations and channels. This grid transformation was performed to accommodate deep learning models that consider channel locations. PSD divides EEG signals into the same four frequency bands as DE and calculates PSD for each band. Extracted PSD features are log-transformed and transformed into a 9 × 9 grid array as input for the model. We also standardized EEG raw signals for deep learning models with feature extraction layers.

Network Configuration and Parameter Settings. In all stages of the experiments, the loss function was Cross-Entropy Loss, the optimization algorithm was Adam optimizer, and the learning scheduler applied was Cosine Annealing Warmup Restarts. The batch size was 64. The maximum learning rate (LR) was set to 10^{-4} for CCNN and 10^{-3} for TSception and DGCNN. The dropout rate was set to 0.5 for CCNN and TSception. Model parameters were determined when the validation loss converged during training processes.

4 Results

Table 1 reports the comparative performance of the proposed framework on the three datasets. Intuitively, all methods combined with ours have an average

accuracy improvement of 2.98%, with the most significant gain of 6.7% over the methods without the proposed OOD detection. This result confirms that our approach is universal and can be applied to various deep-learning methods for building a subject-independent ID model, alleviating the *Concept Shift* problem. We analyze and visualize neural activities with the results from the CCNN-DE method in Sect. 5.4 to further support our claim since they recorded the highest accuracy and AUROC across all datasets. As reported in Table 2, the proposed ODM model consistently recognized different emotions represented by DE features well in both sensitivity and specificity. This result indicates that the better the ODM discriminates data samples between High and Low Groups, the higher

Table 1. Classification Accuracy/AUROC of Baseline, RP, and Ours. Three thresholds (T) were performed per model, and Removal Rate (RR) is the percentage of data removed.

			SEED		SEED-IV		GAMEEMO		
Model		*RR*	Emotion(3)	*RR*	Emotion(4)	*RR*	Valence(9)	*RR*	Arousal(9)
TSception	Baseline	-	91.33 / 93.48	-	79.18 / 85.86	-	87.94 / 93.26	-	88.96 / 93.84
	$RP = 1.5$	*45.7*	**93.47 / 95.10**	*30.1*	**84.71 / 89.80**	*13.7*	89.22 / 93.93	*13.7*	88.99 / 93.94
	$RP = 2.5$	*10.6*	93.10 / 94.81	*11.6*	83.64 / 89.07	*1.5*	88.78 / 93.67	*1.5*	89.48 / 94.10
	$T = .65$	*5.9*	91.91 / 93.93	*25.2*	80.33 / 86.63	*9.3*	90.14 / 94.30	*8.1*	90.73 / 94.85
	$T = .70$	*7.7*	92.86 / 94.64	*30.7*	80.16 / 86.66	*12.0*	88.27 / 93.45	*10.7*	**91.03 / 94.94**
	$T = .80$	*9.7*	92.55 / 94.40	*36.7*	81.07 / 87.25	*18.3*	**90.82 / 94.76**	*15.8*	90.27 / 94.57
CCNN -PSD	Baseline	-	84.92 / 88.66	-	67.76 / 78.66	-	73.74 / 84.67	-	74.01 / 85.46
	$RP = 1.5$	*45.7*	84.18 / 88.11	*30.1*	67.11 / 78.09	*13.7*	76.83 / 86.78	*13.7*	74.02 / 85.61
	$RP = 2.5$	*10.6*	84.36 / 88.26	*11.6*	66.96 / 78.14	*1.5*	73.48 / 84.89	*1.5*	73.46 / 85.40
	$T = .50$	*2.0*	85.68 / 89.23	*15.5*	70.56 / 80.33	*9.5*	75.85 / 85.64	*12.9*	75.61 / 86.33
	$T = .55$	*5.5*	84.82 / 88.60	*23.3*	71.47 / 80.97	*14.4*	76.56 / 86.26	*19.0*	76.44 / 86.85
	$T = .60$	*9.2*	**86.82 / 90.03**	*30.2*	**72.07 / 81.30**	*19.7*	**77.35 / 87.12**	*24.6*	**76.58 / 86.99**
CCNN -DE	Baseline	-	97.48 / 98.11	-	92.25 / 94.86	-	91.18 / 94.69	-	90.40 / 94.42
	$RP = 1.5$	*45.7*	96.83 / 97.62	*30.1*	92.72 / 95.14	*13.7*	92.35 / 95.50	*13.7*	90.73 / 94.81
	$RP = 2.5$	*10.6*	97.66 / 98.24	*11.6*	93.60 / 95.75	*1.5*	91.31 / 94.92	*1.5*	91.50 / 95.20
	$T = .85$	*7.3*	97.92 / 98.44	*13.1*	93.80 / 95.89	*20.1*	94.48 / 96.60	*14.7*	93.17 / 96.08
	$T = .90$	*9.3*	98.11 / 98.57	*16.0*	93.89 / 96.03	*24.9*	95.57 / 97.30	*18.3*	93.76 / 96.63
	$T = .95$	*12.2*	**98.39 / 98.79**	*21.1*	**94.72 / 96.50**	*32.4*	**96.14 / 97.72**	*23.9*	**94.48 / 97.03**
DGCNN -PSD	Baseline	-	83.60 / 87.66	-	65.94 / 77.31	-	70.86 / 83.40	-	69.75 / 82.96
	$RP = 1.5$	*45.7*	82.83 / 87.10	*30.1*	65.60 / 77.04	*13.7*	72.14 / 83.31	*13.7*	72.27 / 84.58
	$RP = 2.5$	*10.6*	82.77 / 87.06	*11.6*	66.50 / 77.65	*1.5*	70.92 / 83.57	*1.5*	70.03 / 83.26
	$T = .50$	*19.9*	87.37 / 90.38	*16.8*	68.86 / 79.35	*15.0*	73.31 / 84.46	*11.6*	73.48 / 85.13
	$T = .55$	*25.0*	87.76 / 90.70	*24.5*	69.43 / 79.78	*21.7*	75.29 / 86.02	*17.8*	**74.54 / 85.88**
	$T = .60$	*30.4*	**88.97 / 91.43**	*31.6*	**70.24 / 80.08**	*27.7*	**76.08 / 86.13**	*23.5*	**74.54** / 85.77
DGCNN -DE	Baseline	-	96.09 / 97.06	-	88.69 / 92.52	-	86.76 / 92.49	-	87.94 / 93.34
	$RP = 1.5$	*45.7*	95.59 / 96.68	*30.1*	89.08 / 92.71	*13.7*	89.50 / 93.93	*13.7*	88.44 / 93.40
	$RP = 2.5$	*10.6*	96.53 / 97.26	*11.6*	89.84 / 93.27	*1.5*	86.38 / 91.97	*1.5*	87.65 / 92.99
	$T = .85$	*7.2*	97.17 / 97.87	*17.7*	91.58 / 94.40	*18.0*	91.55 / 95.05	*18.7*	90.70 / 94.90
	$T = .90$	*9.0*	97.18 / 97.88	*21.9*	92.10 / 94.74	*21.6*	91.35 / 94.72	*23.6*	91.83 / 95.61
	$T = .95$	*12.2*	**97.72 / 98.28**	*28.9*	**92.73 / 95.15**	*27.8*	**93.43 / 95.80**	*29.7*	**92.79 / 95.98**

Table 2. OOD Detection Model (ODM) Performance: Sensitivity/Specificity for High and Low Groups.

Methods		SEED	SEED-IV	GAMEEMO	
				Valence	Arousal
TSception		0.92 / 0.12	0.65 / 0.39	0.83 / 0.36	0.92 / 0.21
CCNN	PSD	0.93 / 0.14	0.74 / 0.43	0.85 / 0.36	0.79 / 0.41
	DE	0.95 / 0.29	0.89 / 0.39	0.75 / 0.67	0.81 / 0.52
DGCNN	PSD	0.73 / 0.42	0.73 / 0.43	0.77 / 0.43	0.84 / 0.32
	DE	0.93 / 0.25	0.78 / 0.50	0.79 / 0.57	0.72 / 0.52

the gain in classification accuracy. RP with TSception outperformed our model in SEED and SEED-IV, showing a decrease of 0.7% and 3.1%, respectively. This observation partially supports the efficacy of detecting outliers at an abnormal distance from the X distribution without considering the Y distribution. However, the model had decrements in performance by about 2.82% compared with ours when experimented on the GAMEEMO dataset. The RP model could not prevail over the intra- and inter-subject variability in light of the imbalanced data distribution.

The results on GAMEEMO demonstrate the superiority of the proposed framework on unbalanced datasets. While T increased, the proposed framework enabled all models to have consistent increments up to 2.7% across all datasets, whereas the RP model did not. This result implies that our strategy provides more efficient structures to detect abnormal samples that may cause *Concept Shift* and prevent models from collapsing due to low SNR by removing the samples.

5 Discussion

5.1 Effect of the Hyper-Parameters.

Since the parameter T is a model's confidence score, its variables determine the amount of outliers. As shown in Fig. 4, as many outliers were detected and removed, the model exhibits a steeper curve than the RP model. The proposed model takes less training time for the same number of iterations and reaches better accuracy faster[1]. As reported in Table 1, whereas the RP method limits improvement of the accuracies to less than 1%, our framework enables an increase in performance up to 2%. This observation implies that our framework strategy provides more efficient structures to achieve a high signal-to-noise ratio.

[1] Xeon-Gold 6330 CPU, 256GB RAM, A6000 GPU.

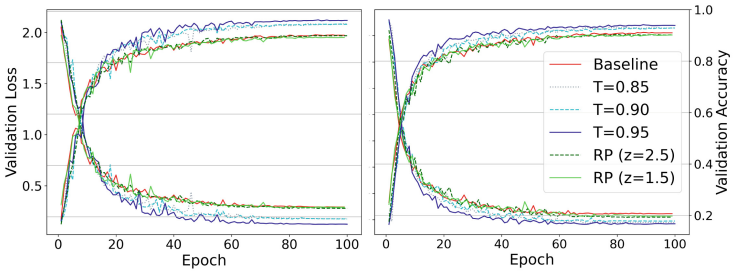


Fig. 4. Convergence curves of validation loss and accuracies on the GAMEEMO-(a) Valence and -(b) Arousal datasets during training.

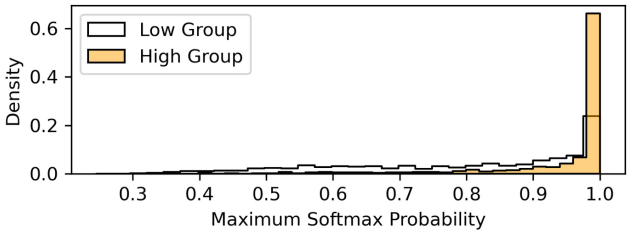


Fig. 5. Distribution of MSP corresponding to High and Low Groups in GAMEEMO-valence dataset.

5.2 OOD Detection Performance and MSP Distribution.

In classification problems, ID and OOD classes are generally separated. ID data belongs to specific class labels, while OOD data does not. This separation helps the model maintain its classification performance on ID while accurately identifying OOD data. However, in this study, ID and OOD are set within the same dataset, leading to overlapping distributions. Consequently, our ODM's performance is lower than typical OOD detection benchmarks due to the overlapping MSP distribution seen in Fig. 5. For the same reason, increasing T also increases the OOD Removal Rate (RR).

5.3 ODM's Impact on Individual and Group-Level Accuracy

Figure 6 depicts the ratio of ID and OOD data (on the left) and the corresponding test results (on the right) for each subject in the GAMEEMO dataset, categorized by arousal (a) and valence (b) labels, using the CCNN-DE method (the left column corresponds to $T = 0.95$). As expected, individuals in the Low Group generally experienced a decrease in accuracy, likely due to the reduced data sample resulting from the exclusion of Low Group data during the ODM construction. However, there were notable exceptions. For example, subject 8 in GAMEEMO-arousal achieved higher accuracy than the baseline, despite having

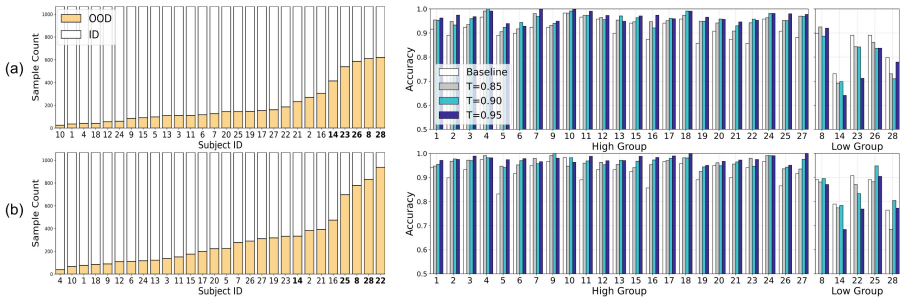


Fig. 6. Left shows the ratio of ID/OOD samples per subject (bold indicates Low Group) and Right shows subject-group accuracies for Baseline and Ours in GAMEEMO-(a) Arousal and -(b) Valence datasets.

the second-largest amount of OOD data removal (Fig. 6 (a)). Similarly, subjects 8, 25, and 28 in GAMEEMO-valence also exhibited increased accuracy, even though they were assigned to the Low Group (Fig. 6 (b)). On the other hand, subjects 16, 21, and 2 in GAMEEMO-valence were in the High Group but had many samples removed. Surprisingly, these individuals showed a noteworthy increase in individual accuracy. These observations suggest that our ODM effectively captured the subject-invariant characteristics of the High Group and contributed to building a robust classification model by eliminating inappropriate samples.

5.4 Qualitative and Statistical Analysis on Neural Patterns

To gain deeper insights into the OOD characteristics under *Concept Shift*, we provide neuroscientific mechanisms by employing EEG data analysis and visualization. Through these analyses, we aim to discern and elucidate the distinctions in neural patterns between the ID and OOD data groups, shedding light on the underlying essence of OOD data and its impact on *Concept Shift*. Throughout this section, we primarily leverage data from the CCNN model with DE features ($T = 0.95$) as it consistently yields the best results across all datasets.

Figure 7 depicts the comparative topographic map of DE between the ID and OOD data, specifically for the affective label with the highest percentage of OOD data in the GAMEEMO dataset. DE values from all ID and OOD data channels within the same frequency band were normalized to fall within the range of (0, 1). Statistical analyses were also conducted through t-tests to assess the differences in DE values between these two groups across all channels. The resulting p-values are displayed in the last row of each subfigure. Within the GAMEEMO dataset, the arousal label of 4 (depicted in Fig. 7 (a)) and the valence label of 1 (shown in Fig. 7 (b)) exhibit the highest OOD percentages at 24% and 49%, respectively. In the SEED (Fig. 7 (c) and SEED-IV (Fig. 7 (d)) datasets, the highest OOD percentages are observed in the affective labels "Positive"

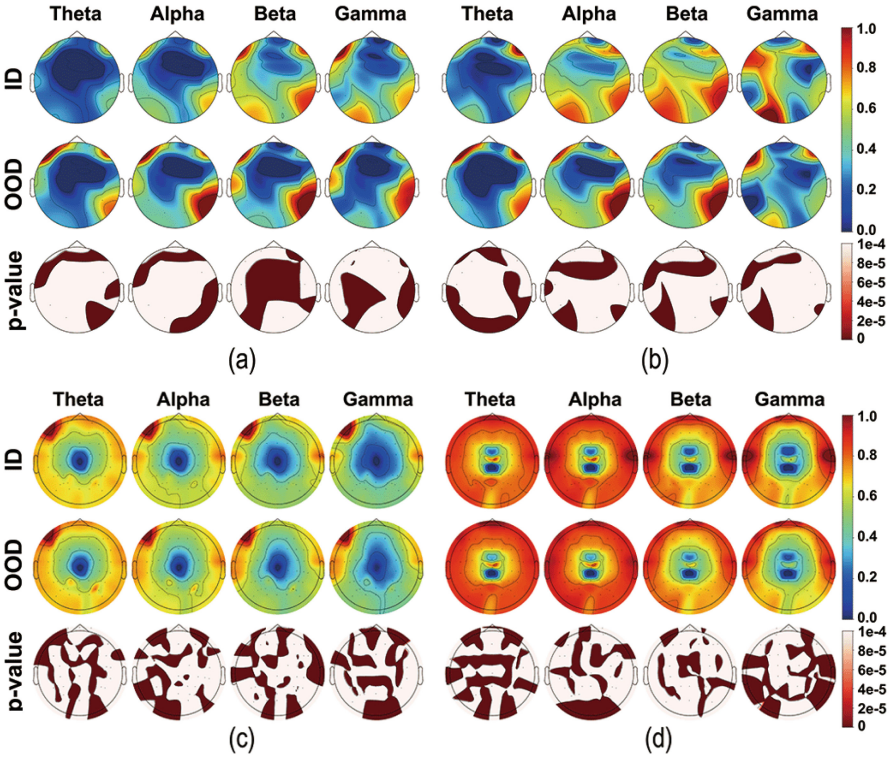


Fig. 7. Topographic map of differential entropy values in four frequency bands for ID and OOD data of (a) GAMEEMO dataset with arousal label of 4, (b) GAMEEMO dataset with valence label of 1, (c) SEED dataset with "Positive" label, and (d) SEED-IV dataset with "Sad" label.

(9.1%) and "Sad" (17%), respectively, but with only marginal differences in OOD percentages among the other classes.

For the GAMEEMO dataset, an arousal label of 4 on a scale of 1 to 9 typically represents a neutral affective state characterized by normal activity across all frequency bands. However, upon closer examination, OOD data exhibit notable distinctions. These differences manifest as significantly heightened activity in the frontal region, particularly in the theta band, as well as in both the frontal and occipital regions, along with the right temporal region in the alpha band. Notably, both the theta and alpha band dominances are typically associated with relaxed or low arousal states [19], presenting a perplexing contrast with the expected neutral state. In contrast, ID data showcase elevated beta and gamma activity, indicating a higher level of arousal compared to OOD data. These intriguing observations suggest that OOD data exhibit a lower level of arousal than the ID data, supporting the hypothesis that challenges posed by

Concept Shift may, in part, arise from inaccuracies in self-assessments of emotional labels.

For the valence label of 1 in the GAMEEMO dataset (Fig. 7 (b)), OOD data consistently exhibit heightened activity in the left frontal region across all frequency bands, while ID data demonstrate increased activity in the right frontal region, particularly in the alpha and beta bands. This intriguing contrast aligns with the notion that negative emotions (low valence) tend to manifest as higher activity in the right hemisphere according to the well-established hemispheric valence theory [21]. Additionally, OOD data reveal significantly higher activity in the right temporal and parietal areas across the theta, alpha, and beta bands. In contrast, ID data display elevated activity in the left temporal and parietal regions, particularly in the alpha, beta, and gamma bands.

In Fig. 7 (c), a similar pattern in the distribution of DE across the scalp is observed in the "Positive" data of the SEED dataset for both ID and OOD data. Positive emotions are typically associated with higher activity in the left hemisphere [21], a pattern evident across all frequency bands in both ID and OOD groups. However, results from t-tests reveal significant differences between the two groups in several scalp regions across all frequency bands. These findings suggest that while the neural patterns in both ID and OOD data align with the expected pattern for "Positive" emotions, the OOD data exhibit distinct details that significantly differ from the ID data. One plausible explanation for this phenomenon is that the affective labels in the SEED dataset primarily account for the valence aspect of affective states, encompassing only "Negative", "Neutral", and "Positive" emotions, while omitting considerations of arousal levels. Consequently, the OOD data within the 'Positive' class may represent instances with varying arousal levels compared to the norm established by the ID data.

The distributions of DE across the scalp in the "Sad" data of SEED-IV dataset are shown in Fig. 7 (d). "Sad" emotion, which encompass both medium-to-low arousal and low valence, are theoretically associated with high activity in the lower frequency bands and a greater involvement of the right hemisphere [6]. From the figure, higher activity in the right temporal lobe can be observed for both ID and OOD data. OOD data consistently exhibit significantly higher activity in the theta and alpha band than the ID data, suggesting a potential tendency toward lower arousal levels. In contrast, ID data showcase heightened gamma band activity, indicating a higher level of arousal [13]. This situation parallels the issue observed in the SEED dataset, where the discrete emotion approach [5] is used to describe affective states. While the SEED-IV dataset includes the label "fear", which theoretically associates with high arousal and low valence, it lacks affective labels that pertain to a wide range of emotional spectrum.

As a culmination of our analysis, two critical insights emerge regarding the OOD data challenge within the *Concept Shift* problem. Firstly, our observations indicate that OOD data might arise due to inaccuracies in self-assessments of affective states. These findings underscore that, despite sharing the same affective state label, OOD data can exhibit varying degrees of specific neural

activation, highlighting the nuanced nature of emotional experiences. Secondly, the challenge is exacerbated by the use of the discrete emotion approach to label affective states, particularly when the number of labels is limited. This concern is evident in both the SEED and SEED-IV datasets, where affective state labels fail to encompass the full spectrum of human emotions. Consequently, neural and data patterns within the same class may vary significantly, reflecting the intricate and multifaceted nature of the cognitive processes underlying emotions. These insights reveal the complexity of addressing the OOD data challenge and emphasize the need for more precise and comprehensive approaches to affective state labeling in the field of affective computing research.

6 Conclusion

Our proposed framework improved classification accuracy in public datasets, addressing *Concept Shift* issues due to unreliable self-assessments in EEG-based emotion classification through OOD detection. We introduced a unique solution that leverages overlooked inter-subject variability by categorizing subjects into High and Low Groups based on the Top-K results of subject-dependent. This approach is beneficial for datasets where human labeling is unreliable due to inherent ability in individual cognition and perception. Our findings emphasize the need to consider OOD during data collection if its distribution can differ by user characteristics, advocating for recording latent variables to refine data and improve model performance. While we focused on optimizing the model for the majority of subjects, we have laid the groundwork for developing tailored enhancement plans for the few low-performing subjects. This approach ensures that we can address the specific challenges faced by these individuals in future work.

References

1. Alakus, T.B., Gonen, M., Turkoglu, I.: Database for an emotion recognition system based on EEG signals and various computer games-GAMEEMO. Biomed. Signal Process. Control **60**, 101951 (2020)
2. Alarcão, S.M., Fonseca, M.J.: Emotions recognition using EEG signals: a survey. IEEE Trans. Affect. Comput. **10**(3), 374–393 (2019)
3. Barthélemy, Q., Mayaud, L., Ojeda, D., Congedo, M.: The Riemannian potato field: a tool for online signal quality index of EEG. IEEE Trans. Neural Syst. Rehabil. Eng. **27**(2), 244–255 (2019)
4. Can, Y.S., Mahesh, B., André, E.: Approaches, applications, and challenges in physiological emotion recognitiona tutorial overview. In: Proceedings of the IEEE (2023)
5. Christie, I.C., Friedman, B.H.: Autonomic specificity of discrete emotion and dimensions of affective space: a multivariate approach. Int. J. Psychophysiol. **51**(2), 143–153 (2004)

6. Crawford, H.J., Clarke, S.W., Kitner-Triolo, M.: Self-generated happy and sad emotions in low and highly hypnotizable persons during waking and hypnosis: laterality and regional EEG activity differences. Int. J. Psychophysiol. **24**(3), 239–266 (1996)
7. Ding, Y., Robinson, N., Zhang, S., Zeng, Q., Guan, C.: Tsception: capturing temporal dynamics and spatial asymmetry from EEG for emotion recognition. IEEE Trans. Affect. Comput. (2022)
8. Fort, S., Ren, J., Lakshminarayanan, B.: Exploring the limits of out-of-distribution detection. Adv. Neural Inf. Process. Syst. (NeurIPS) **34**, 7068–7081 (2021)
9. Hendrycks, D., Mazeika, M., Dietterich, T.: Deep anomaly detection with outlier exposure. In: International Conference on Learning Representation (ICLR) (2019)
10. Hoemann, K., Lee, Y., Kuppens, P., Gendron, M., Boyd, R.L.: Emotional granularity is associated with daily experiential diversity. Affect. Sci. **4**(2), 291–306 (2023)
11. Huang, R., Geng, A., Li, Y.: On the importance of gradients for detecting distributional shifts in the wild. Adv. Neural Inf. Process. Syst. (NeurIPS) **34**, 677–689 (2021)
12. Jiang, X., et al.: Detecting out-of-distribution data through in-distribution class prior. In: International Conference on Machine Learning (ICML), vol. 202, pp. 15067–15088 (2023)
13. Keil, A., Müller, M.M., Gruber, T., Wienbruch, C., Stolarova, M., Elbert, T.: Effects of emotional arousal in the cerebral hemispheres: a study of oscillatory brain activity and event-related potentials. Clin. Neurophysiol. **112**(11), 2057–2068 (2001)
14. Li, X., et al.: EEG based emotion recognition: a tutorial and review. ACM Comput. Surv. **55**(4), 1–57 (2022)
15. Liu, W., Wang, X., Owens, J., Li, Y.: Energy-based out-of-distribution detection. Adv. Neural Inf. Process. Syst. (NeurIPS) **33**, 21464–21475 (2020)
16. Morris, J.D.: Observations: SAM: the self-assessment manikin; an efficient cross-cultural measurement of emotional response. J. Advert. Res. **35**(6), 63–68 (1995)
17. Muller, M., et al.: Designing ground truth and the social life of labels. In: CHI Conference on Human Factors in Computing Systems (CHI) (2021)
18. Nguyen, A., Yosinski, J., Clune, J.: Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 427–436 (2015)
19. Reuderink, B., Mühl, C., Poel, M.: Valence, arousal and dominance in the EEG during game play. Int. J. Autonom. Adapt. Commun. Syst. **6**(1), 45–62 (2013)
20. Saganowski, S., Perz, B., Polak, A.G., Kazienko, P.: Emotion recognition for everyday life using physiological signals from wearables: a systematic literature review. IEEE Trans. Affect. Comput. **14**(3), 1876–1897 (2023)
21. Silberman, E.K., Weingartner, H.: Hemispheric lateralization of functions related to emotion. Brain Cogn. **5**(3), 322–353 (1986)
22. Song, T., Zheng, W., Song, P., Cui, Z.: EEG emotion recognition using dynamical graph convolutional neural networks. IEEE Trans. Affect. Comput. **11**(3), 532–541 (2020)
23. Tian, L., et al.: Recognizing induced emotions of movie audiences: Are induced and perceived emotions the same? In: International Conference on Affective Computing and Intelligent Interaction (ACII), pp. 28–35 (2017)
24. Ward, M., Gruppen, L., Regehr, G.: Measuring self-assessment: current state of the art. Adv. Health Sci. Educ. **7**, 63–80 (2002)

25. Yang, J., Zhou, K., Li, Y., Liu, Z.: Generalized out-of-distribution detection: a survey. arXiv preprint arXiv:2110.11334 (2021)
26. Yang, Y., Wu, Q., Fu, Y., Chen, X.: Continuous convolutional neural network with 3D input for EEG-based emotion recognition. In: International Conference on Neural Information Processing(ICONIP), pp. 433–443. Springer (2018)
27. Zheng, W.L., Liu, W., Lu, Y., Lu, B.L., Cichocki, A.: Emotionmeter: a multimodal framework for recognizing human emotions. IEEE Trans. Cybern. **49**(3), 1110–1122 (2018)
28. Zheng, W.L., Lu, B.L.: Investigating critical frequency bands and channels for EEG-based emotion recognition with deep neural networks. IEEE Trans. Auton. Ment. Dev. **7**(3), 162–175 (2015)

A Trainable Feature Extractor Module for Deep Neural Networks and Scanpath Classification

Wolfgang Fuhl(✉)

University Tübingen, Sand 14, 72076 Tübingen, Germany
wolfgang.fuhl@uni-tuebingen.de

Abstract. Scanpath classification is an area in eye tracking research with possible applications in medicine, manufacturing as well as training systems for students in various domains. In this paper we propose a trainable feature extraction module for deep neural networks. The purpose of this module is to transform a scanpath into a feature vector which is directly useable for the deep neural network architecture. Based on the backpropagated error of the deep neural network, the feature extraction module adapts its parameters to improve the classification performance. Therefore, our feature extraction module is jointly trainable with the deep neural network. The motivation to this feature extraction module is based on classical histogram-based approaches which usually compute distributions over a scanpath. We evaluated our module on three public datasets and compared it to the state of the art approaches.

Keywords: Neural Network · Deep Neural Network · Modul · Saccade · Classification · Scanpath classification

1 Introduction

Our eye movements reveal more than just what we see. They also show how our brain and senses work together. By looking at the sequence and duration of eye fixations and jumps, called the scanpath, we can learn how we process information. In different fields, eye tracking studies have discovered patterns in eye movements. These patterns can distinguish between different types of people (e.g., beginners and experts), or different situations, such as the task given to a person. For example in art, eye movement differences have been observed between professional and novice art viewers for both realistic and abstract art [46]. Moreover, top-down expectations and bottom-up visual features can influence the eye movements on artworks [33,35]. Similarly in the medical field, scanpath differences can indicate the professional and the treatment factors. Scanpath differences between beginners and experts have been reported in microneurosurgeons [14,31] and radiologists [22,34]. It was also found that dental students who took a specific radiography training course could be correctly identified from their scanpaths.

A. Antonacopoulos et al. (Eds.): ICPR 2024, LNCS 15313, pp. 292–304, 2025.
https://doi.org/10.1007/978-3-031-78201-5_19

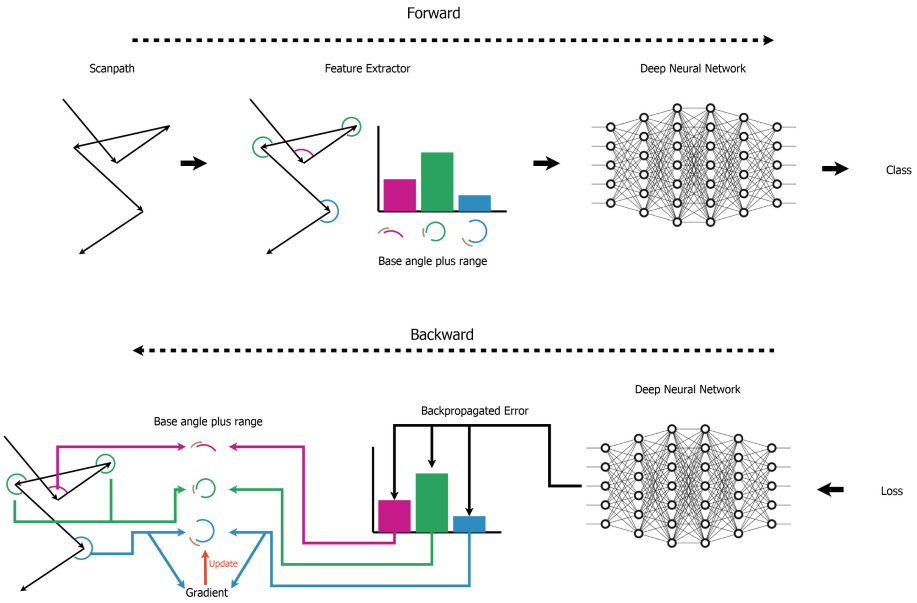


Fig. 1. The forward and backward pass of the proposed feature extractor module.

Regarding the treatment factor, eye movement differences from healthy controls have been observed in both patients with schizophrenia [26] and autism spectrum disorder [36,42]. Therefore, scanpaths can potentially be used for more precise training, diagnosis, and treatment methods. In driving, scanpaths have been applied to reliably assess safe or unsafe driving in people with visual impairments [31]. Moreover, they can be utilized in driver assistance systems to signal the take-over readiness [41] or cognitive load [30], and fatigue [44]. Notably, most of the studies above focus on finding statistically significant differences in single scanpath measures. Hence, there is a large and ever increasing body of scanpath comparison and classification methodology: From simple statistics to state-of-the-art machine learning [7].

In this paper we propose a novel deep neural network layer for scanpath classification which is inspired by the approach from [19]. Our approach is not dependent on generated or predefined areas of interest and works on the gaze samples directly. We reformulated the angle and angle range approach from [19] in a way that it is trainable by the backpropagation algorithm. This means that our novel layer for deep neural networks is jointly trainable and works as a feature extraction module in front of the classification part, which is usually a fully connected stage.

In short, our contributions are:

- A novel feature extraction layer for deep neural networks.
- Integration of the angle and angle range approach into the backpropagation algorithm.

– Evaluations on multiple public datasets and comparison to the state of the art approaches.

2 Related Work

In the 1990 s, the initial proposal of automated metrics took place [5]. Since then, there has been a significant evolution in the methodology for automated scanpath comparison [2]. Recently, machine-learning based approaches have emerged, showing impressive results [10,16,27,47]. These approaches are capable of distinguishing relevant eye movement patterns from high levels of noise. However, there is still an ongoing debate regarding how to efficiently encode eye movement trajectories for machine learning purposes. Some algorithms heavily rely on time-aggregated features or complete-sequence alignment, while others focus on gaze transitions (i.e., the shift of gaze between two targets, also known as a saccade) as a popular feature [6,11,12,27]. This strategy allows for the modeling of cognitive associations between gaze targets. Hidden Markov Models (HMMs) remained long the most common approach [9,24], but there are also other methods that extend these patterns to span multiple subsequent fixations and saccades. Identifying patterns in longer sequences is particularly important, as these patterns can be more specific to a task or subject group, making them highly useful for classification [31].

A CNN related to autism spectrum disorder used fixation maps from the entire scanpath as input [15]. Multiple deep learning models for schizophrenia and task classification also used heatmaps as input [29,43]. To include temporal information, CNN-LSTM networks classified autism spectrum disorder using scanpath-based patches from a saliency map [39]. [15] used gaze snapshots based on group attention at time intervals as input for an autoencoder. These attention map methods often apply techniques like Gaussian blurring to the raw data. Other methods use different scanpath image creations for classifications, such as Markov random fields [45], object detection based [40], input image patch based [8], and principle component based [32].

Creating images from the raw scanpath data is another representation option. This approach preserves potentially relevant information for the model that preprocessing could remove. [38] used scanpath images from the raw gaze for the entire duration and for five second intervals as input for an RNN to classify confusion. [1] used scanpath images by connecting saccades and weighting them by the fixation densities with a CNN. A generative model for scanpath classification that converted gaze data into emojis was proposed in [17]. It encodes gaze data as a compact image with the red, green, and blue channels representing the spatial, temporal, and connectivity which is than fed to a deep neural network. [3,4,7] explored different scanpath representations for classification. They used temporal coloring for saccade velocities or symbols for different fixation durations.

3 Method

Figure 1 shows the workflow of the proposed approach, which is inspired by the random ferns used in [19]. In the forward pass, our layer gets the entire sequence and checks a series of inter-sample angles and angle ranges. Based on these checks, a histogram index is selected and increased. Each bin in this histogram corresponds to an angle and an angle range (Fig. 1 central histogram). If the angle between two samples falls into this angle range, centered around the base angle, the bin is increased. The final histograms are given to a classifier, which is, in our case, a residual neural network with convolutions. This part of our approach is similar to what was done in the original paper with balanced decision trees, also known as random ferns [19] without the residual network as a classifier. The interesting part and the main contribution of our method is the backward pass. Here, we propagate the gradient back to each histogram bin. Based on the sign of the gradient, we either increase or decrease the angle range. In the following, we will describe our module and its integration into the backpropagation algorithm in detail.

Algorithm 1. The algorithmic description of the forward pass for the proposed module. First, we set the histograms to zero. Afterward, all angle and angle range checks are applied to the entire sequence. Based on these checks, an index for the current histogram that belongs to a set of angles and angle ranges is computed. With this index, a histogram bin is increased, and in the end, each histogram is normalized.

procedure ForwardPass($sequence, anglesets, histograms$)
  $histograms = 0$
  for all $set_i \in anglesets$ **do**
    for $seq_j = 0; seq_j < size(sequence - size(set_i)); seq_j++$ **do**
      $index = 0$
      for all $angleANDrange_k \in set_i$ **do**
        if $angleANDrange_k(sequence(seq_j + k))$ **then** ▷ 1 if in range
          $index += 2^k$
      $histograms[index(set_i)][index] += 1$
    for $histo_i = 0; seq_j < size(set_i); seq_j++$ **do** ▷ Normalize histogram
      $histograms[index(set_i)][histo_i] = \frac{histograms[index(set_i)][histo_i]}{\sum_{l=1}^{size(set_i)} histograms[index(set_i)][l]}$

Algorithm 1 describes the forward pass of our approach in the backpropagation algorithm. Before we can use our layer in a neural network we have to specify two parameters, one is the amount of angle sets which is the amount of sequences consisting of angles and angle ranges we want to use. The second parameter is the length of such a sequence. A simplified illustration can be found in Fig. 1. In this illustration, the sequence length would be one, which means that each angle and angle range has its own bin in the histogram. After we set up those two parameters, we evaluate each angle and angle range sequence on the

given samples from an eye tracking recording and compute the histograms as described in Algorithm 1. The filled histograms are normalized and given to a neural network for further processing.

Algorithm 2. The backward pass of our approach is described as an algorithm. In the first part, we have to do the forward pass again. With the computed indexes, we can access the gradients corresponding to different angle and angle range sets. Next, we need to compute which angle and angle range check are evaluated positive. For all positive angle and angle range checks, we sum up the corresponding gradients without updating them directly. In the last step, we update all angle ranges according to the cumulated gradients. We do not update them directly since this would change the angle and angle range checks, which would invalidate our gradient computation. In the real implementation, we do not need to compute the indexes since they are stored with the forward pass. In addition, we also do not need to evaluate which angle and angle range check evaluated positive, since this is already known based on the histogram index. This means that we described the backward pass in a way it can be computed and that can be understood more easily. The real implementation therefore differs from the algorithm to save resources and training time.

procedure BACKWARDPASS($sequence, anglesets, gradient, learningrate$)
  for all $set_i \in anglesets$ **do**
    for $seq_j = 0; seq_j < size(sequence - size(set_i)); seq_j++$ **do**
      $index = 0$
      for all $angleANDrange_k \in set_i$ **do**
        if $angleANDrange_k(sequence(seq_j + k))$ **then** ▷ 1 if in range
          $index += 2^k$
      for all $angleANDrange_k \in set_i$ **do**
        if $angleANDrange_k(sequence(seq_j + k))$ **then** ▷ 1 if in range
          $angleANDrange_k.rangeUpdate += gradient[i][index]$
            ▷ Computation of the cumulative update of the angle range
  for all $set_i \in anglesets$ **do**
    for all $set_i \in anglesets$ **do**
      for all $angleANDrange_k \in set_i$ **do**
        $angleANDrange_k.range += angleANDrange_k.rangeUpdate * learningrate$
          ▷ Applying the cumulative update to the angle range

Algorithm 2 describes the backward pass of our approach. For simplification, we added the parts of the forward pass into it so it can be fully understood. In the real implementation, we store the result of the forward pass, and we also know based on the histogram bin which angle and angle range evaluated to one due to the way we set them up in our memory. The input to the neural network, our histograms, receive the back propagated error from the network. With the error or gradient assigned to each bin in our histograms, we can compute which angle and which angle range has evaluated to one. The corresponding angle ranges are

then adjusted based on the gradient. If the gradient is negative, we reduce the angle range and if it is positive, we increase the angle range. Since multiple angle ranges participated on multiple places over a sequence, we accumulate the gradient first. This is indicated by ".rangeUpdate" in Algorithm 2.

4 Evaluation

In this section we first describe the used public datasets and how we performed the training, validation and test splits. Afterward, the training parameters of our approach and the configuration of the other approaches is described. The last part in this section presents and discusses our results.

4.1 Datasets

Gaze [13]: A data set with eye tracking data on moving scenes. The data was collected using an SR Research EyeLink II eye tracker with 250 Hz. For our experiment, we used the data given for static images where each static image of a video was treated as the same image. Moreover, we omitted subject V01 since there was only one recording available. Hence, we used the eye tracking data of 10 subjects on 9 images for our experiment with an average recording duration of 2 s. The training and test split was done using 50% for the training and 50% for the testing with a random selection. For the stimulus classification we made sure that no subject is shared between the training and testing set and for the subject classification we did the same based on the stimulus.

WherePeopleLook [28]: An eye tracking dataset that focuses on integrating top-down features into the generation of saliency maps. This dataset comprises 1003 static images, each accompanied by eye tracking data from 15 subjects. The eye tracking data was collected for an average recording length of 3 s per image. To conduct our experiment, we divided the dataset into a training set and a testing set, ensuring a balanced split of 50% for each. We took great care to ensure that no subject appeared in both the training and testing sets for stimulus classification. Similarly, for subject classification, we made sure that no stimulus overlapped between the training and testing sets. This meticulous approach guarantees the integrity and reliability of our experiment results.

DOVES [37]: An extensive eye tracking dataset consisting of data from 29 subjects recorded on 101 natural images. The recordings were conducted using a high-precision dual-Purkinje eye tracker with a sampling rate of 200 Hz. Each recording had an average length of 5 s. Following the approach used in the WherePeopleLook dataset, we split the dataset into training and testing sets, with an equal distribution of 50% for each. To ensure accurate stimulus classification, we took care to avoid any overlap of subjects between the training and testing sets. Similarly, for subject classification, we ensured that no stimulus was shared between the training and testing sets. This rigorous methodology ensures the reliability and validity of our dataset for further analysis and experimentation.

We decided to use those datasets since they have a large amount of available sequences, which is important for neural network based approaches.

4.2 Training and Adaption of the Other Methods

All used CNNs (Convolution neural networks) are ResNet-12 [25] in our evaluation. This concerns all indications of "+ CNN" in Table 2 and our approach. For the training of those networks, we used a soft max classification layer with the stochastic gradient descent optimizer and momentum. The initial learning rate was set to 10^{-3} and reduced to 10^{-4} after 50 epochs. With the learning rate of 10^{-4} we trained additional 50 epochs and used the best model based on the results of the validation set which consists of 20% randomly selected from our training set. For the features HOV and HEAT we computed multiple parameter configurations and selected the best model based on the 20% validation set. For [7] we evaluated all representations and selected the best performing representation based on the results on the 20% validation set. For *RNN [38] and *LSTM [39] we did not find the code online and therefore tried to reproduce the approach as best as we could. For Subsmatch 2.0 we selected the best performing AOI selection approach, and we also tried different classifiers for the features as well as different parameters for the classifiers. For the random fern approach, we evaluated different feature selection and ensemble combination parameters and selected the best performing one. For the *EM Statistics and Auto AOI + statistics we used duration, speed, and acceleration based on the AOIs or the eye movement types which were classified using a velocities and dispersion threshold. Based on duration, speed, and acceleration, we computed the mean, variance, standard deviation, and the confidence intervals. All together was one feature vector. The deep semantic gaze embedding was also reproduced with a ResNet-12. For the Encodji approach, we used the ResNet-12 as classifier as well as discriminator to train the generative adversarial network. The generative adversarial network itself was a U-Net with interconnections. All approaches had the same training, validation and testing data.

4.3 Results

In Table 1 we evaluated different combinations of our two parameters amount of angle sets and the set size. For the evaluation, we used 80% of the training data and 20% of the training data for validation. The reported accuracy is rounded and computed on the validation set. As can be seen, the amount of sequences has a huge impact on the classification accuracy. This is the case since randomly selected angle and angle range sequences can also be useless. With more such sequences, we increase the chance in getting good combinations. For the set size parameter the same is true since larger sequences have a lower probability of a good selection since there are more possible combinations. This means that both parameters are in relation to each other, which means that higher set sizes also require larger amounts of angle sets. Based on our evaluation, we have selected $2^{12} = 4.096$ angle sets and an angle as well as angle range sequence

Table 1. The metric in this table is accuracy rounded to two decimal places. We evaluated different initializations of our approach for the dataset Gaze. We performed the subject and stimulus classification. Angle sets are the amount of all sequences of angles and angle ranges that are randomly initialized and evaluated during the forward pass. Set size is the amount of angles and angle ranges that are in one angle set. For the evaluation we performed a 80% training and 20% validation split on the training data only.

Angle sets	Set size	Dataset Gaze	
		10 Classes Subject	9 Classes Stimulus
2^9	4	54	22
2^9	5	68	31
2^9	6	71	34
2^{10}	4	73	39
2^{10}	5	82	51
2^{10}	6	81	47
2^{11}	4	78	43
2^{11}	5	89	58
2^{11}	6	87	55
2^{12}	4	82	49
2^{12}	5	**91**	**62**
2^{12}	6	88	57

length of 4. For binary decisions, end up in $2^5 = 32$ possible combinatorial outcomes. Therefore, our produced tensor for the neural network has a size of 4.096×32.

Table 2 shows the comparison to other state of the art approaches with the accuracy metric. As can be seen, our approach works similarly well as the random fern approach from which we used the feature extraction approach. This means that the sequence of angles and angle ranges seams to be a good feature for the used datasets or scanpath in general. What can be seen also is that all neural network based approaches, excluding the ones on statistics, worked well on the datasets. The reason for this is possibly that the large amount of available sequences helped the neural networks. The statistical based approaches as well as the features HEAT and HOV worked not so well on the datasets, which is due to the short recording length. Statistics and features like the histogram of oriented gradients or the heatmaps need possibly more data to be discriminative. The worst approaches were the RNN and the LSTM, they suffer the most from the short recording length since they require long sequences for which they are designed. In total, our approach was at least as good as the random ferns. Sometimes our approach outperformed the ferns, and for the stimulus classification in the WherePeopleLook dataset the ferns were slightly better compared to our approach. Therefore, we think our approach to integrate the angle and

Table 2. The used metric is accuracy, which we rounded to two decimal places. We compared our approach with the best parameters from Table 1 with other state of the art approaches. Best results are shown in bold, and * indicates that we have reimplemented the methods as it is described in Sect. 4.2. Evaluated on the testing data.

Dataset	Gaze		WherePeopleLook		DOVES	
Target	Sub	Stim	Sub	Stim	Sub	Stim
Classes	10	9	15	1003	29	101
*Deep semantic gaze embedding [8]	80	38	39	37	15	48
Random Ferns [19]	85	44	**42**	**41**	18	**52**
Subsmatch 2.0 [31]	69	28	27	30	8	30
*EM Statistics [23] + SVM	61	22	21	20	5	28
*EM Statistics [23] + Tree ENS	62	24	25	21	6	31
*EM Statistics [23] + two layer NN	60	24	23	21	6	29
Auto AOI [21] + statistics + SVM	69	27	29	27	9	39
Auto AOI [21] + statistics + Tree ENS	72	29	30	28	10	41
Auto AOI [21] + statistics + two layer NN	71	29	29	28	10	40
Encodji [17]	83	39	36	34	16	48
Encodji input only [17] + CNN	77	31	32	31	11	45
Best from [7] + CNN	78	33	31	33	12	47
HEAT + SVM [20]	74	28	30	29	8	46
HEAT + Tree ENS [20]	76	29	33	31	10	47
HEAT + two layer NN [20]	75	26	30	28	9	44
HOV + SVM [18]	74	31	31	28	10	45
HOV + Tree ENS [18]	75	30	33	30	11	46
HOV + two layer NN [18]	73	27	31	26	9	43
*RNN [38]	70	26	27	22	6	31
*LSTM [39]	67	23	26	20	5	27
Proposed	**87**	**45**	**42**	40	**19**	**52**

angle range feature into deep neural networks or into the backpropagation algorithm was successful. The advantage of our approach vs the Random Ferns are, that our approach can be trained with batches instead of requiring the entire dataset during training like the random ferns do. In addition, the ResNet-12 can be replaced with larger models. As soon as larger datasets for eye tracking are available, it would be possible to use transformers for example.

5 Limitations

We compared our approach on three public datasets with different amounts of subjects and stimuli. While this can be seen as an extensive evaluation, it is not

guaranteed that the results apply for all datasets. One example here are long term recordings in real life which are publicly not available as a dataset, which is also the reason why we did not evaluate on such datasets. Another limitation of our paper are the optimal parameters for the state of the art approaches. We tried to reproduce the methods and approaches as good as we could and also tried to select the best parameters, but it is still possible that there are preprocessing steps or parameter combinations which lead to better result.

6 Conclusion and Outlook

In this paper, we proposed a module for deep neural networks which is based on the angle and angle range approach from [19]. Our main contribution is the integration of the approach into the backpropagation algorithm, which makes it possible to train it jointly with deep neural networks or any other computational graph based approach which requires derivatives for gradient determination. Our approach outperformed most of the state of the art approaches, but it has to be noted that all of our datasets have only a small recording length since long term recordings with many subjects and many sequences of scanpath are not publicly available, which is especially true for medical recordings. The approach with random ferns is very close to our results, and sometimes beats our deep neural network based approach. This means that it is still the case that entropy and information gain, which is used to train decision trees like the random ferns, can outperform deep neural networks as it is common for tabular data. In addition, it is also obvious that especially for the used datasets, the consecutive sample angles form strong features for the classification. Overall we think that our proposed method to integrate the consecutive angle and angle range approach into the backpropagation algorithm was successful since it delivers results close or slightly better than the original approach based on decision trees [19]. Future work should investigate if it is possible to also learn the base angle itself as well as more advanced module for an internal use in deep neural networks. In addition, it could be possible to use it in generative adversarial networks to generate human visual behavior.

7 Potentially Harmful Impacts and Future Societal Risks

The proposed approach is directed into the research area of eye tracking with the purpose to help humans in terms of a supportive diagnosis system or to be part of educational software. Of course, there are many possibilities to use scanpath classification in harmful ways, like the observation and intention prediction of humans. The classification of abnormal behavior could for example reveal if somebody has a sickness like Alzheimer or Autism, which is private information. Scanpath analysis could also be used to train humans for specific harmful tasks or to cheat in competitions like poker for example. We as researchers do not want our knowledge to be used in such areas or for such tasks, but we cannot prevent it from happening.

Acknowledgement. Funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) - 508330921

References

1. Ahmed, Z.A., Jadhav, M.E.: Convolutional neural network for prediction of autism based on eye-tracking scanpaths. Int. J. Psych. Rehabili. **24**(05) (2020)
2. Anderson, N.C., Anderson, F., Kingstone, A., Bischof, W.F.: A comparison of scanpath comparison methods. Behav. Res. Methods **47**, 1377–1392 (2015)
3. Atyabi, A., et al.: Stratification of children with autism spectrum disorder through fusion of temporal information in eye-gaze scan-paths. ACM Trans. Knowl. Discov. Data **17**(2), 1–20 (2023)
4. Bhattacharya, N., Rakshit, S., Gwizdka, J., Kogut, P.: Relevance prediction from eye-movements using semi-interpretable convolutional neural networks. In: Proceedings of the 2020 Conference on Human Information Interaction and Retrieval, pp. 223–233 (2020)
5. Brandt, S.A., Stark, L.W.: Spontaneous eye movements during visual imagery reflect the content of the visual scene. J. Cogn. Neurosci. **9**(1), 27–38 (1997)
6. Burch, M., Kurzhals, K., Kleinhans, N., Weiskopf, D.: Eyemsa: exploring eye movement data with pairwise and multiple sequence alignment. In: Proceedings of the 2018 ACM Symposium on Eye Tracking Research & Applications, pp. 1–5 (2018)
7. Byrne, S.A., Maquiling, V., Reynolds, A.P.F., Polonio, L., Castner, N., Kasneci, E.: Exploring the effects of scanpath feature engineering for supervised image classification models. In: Proceedings of the ACM on Human-Computer Interaction, vol. 7(ETRA), pp. 1–18 (2023)
8. Castner, N., et al.: Deep semantic gaze embedding and scanpath comparison for expertise classification during opt viewing. In: ACM Symposium on Eye Tracking Research and Applications, pp. 1–10 (2020)
9. Coutrot, A., Hsiao, J.H., Chan, A.B.: Scanpath modeling and classification with hidden markov models. Behav. Res. Methods **50**(1), 362–379 (2018)
10. Crabb, D.P., Smith, N.D., Zhu, H.: What's on tv? detecting age-related neurodegenerative eye disease using eye movement scanpaths. Front. Aging Neurosci. **6**, 312 (2014)
11. Cristino, F., Mathôt, S., Theeuwes, J., Gilchrist, I.D.: Scanmatch: a novel method for comparing fixation sequences. Behav. Res. Methods **42**, 692–700 (2010)
12. Dewhurst, R., Foulsham, T., Jarodzka, H., Johansson, R., Holmqvist, K., Nyström, M.: How task demands influence scanpath similarity in a sequential number-search task. Vision. Res. **149**, 9–23 (2018)
13. Dorr, M., Martinetz, T., Gegenfurtner, K.R., Barth, E.: Variability of eye movements when viewing dynamic natural scenes. J. Vis. **10**(10), 28–28 (2010)
14. Eivazi, S., et al.: Gaze behaviour of expert and novice microneurosurgeons differs during observations of tumor removal recordings. In: Proceedings of the Symposium on Eye Tracking Research and Applications, pp. 377–380 (2012)
15. Elbattah, M., Carette, R., Dequen, G., Guérin, J.L., Cilia, F.: Learning clusters in autism spectrum disorder: image-based clustering of eye-tracking scanpaths with deep autoencoder. In: 2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), pp. 1417–1420. IEEE (2019)
16. French, R.M., Glady, Y., Thibaut, J.P.: An evaluation of scanpath-comparison and machine-learning classification algorithms used to study the dynamics of analogy making. Behav. Res. Methods **49**, 1291–1302 (2017)

17. Fuhl, W., et al.: Encodji: encoding gaze data into emoji space for an amusing scanpath classification approach. In: Proceedings of the 11th ACM Symposium on Eye Tracking Research & Applications, pp. 1–4 (2019)
18. Fuhl, W., Castner, N., Kasneci, E.: Histogram of oriented velocities for eye movement detection. In: Proceedings of the Workshop on Modeling Cognitive Processes from Multimodal Data, pp. 1–6 (2018)
19. Fuhl, W., Castner, N., Kübler, T., Lotz, A., Rosenstiel, W., Kasneci, E.: Ferns for area of interest free scanpath classification. In: Proceedings of the 11th ACM Symposium on Eye Tracking Research & Applications, pp. 1–5 (2019)
20. Fuhl, W., Sanamrad, N., Kasneci, E.: The gaze and mouse signal as additional source for user fingerprints in browser applications. arXiv preprint arXiv:2101.03793 (2021)
21. Fuhl, W., et al.: Area of interest adaption using feature importance. In: Proceedings of the 2023 Symposium on Eye Tracking Research and Applications, pp. 1–7 (2023)
22. van der Gijp, A., Webb, E.M., Naeger, D.M.: How radiologists think: understanding fast and slow thought processing and how it can improve our teaching. Acad. Radiol. **24**(6), 768–771 (2017)
23. Goldberg, J.H., Helfman, J.I.: Visual scanpath representation. In: Proceedings of the 2010 Symposium on Eye-Tracking Research & Applications, pp. 203–210 (2010)
24. Hacisalihzade, S.S., Stark, L.W., Allen, J.S.: Visual perception and sequences of eye movement fixations: a stochastic modeling approach. IEEE Trans. Syst. Man Cybern. **22**(3), 474–481 (1992)
25. 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)
26. Hooker, C., Park, S.: You must be looking at me: the nature of gaze perception in schizophrenia patients. Cogn. Neuropsychiatry **10**(5), 327–345 (2005)
27. Hoppe, S., Loetscher, T., Morey, S.A., Bulling, A.: Eye movements during everyday behavior predict personality traits. Front. Hum. Neurosci. 105 (2018)
28. Judd, T., Ehinger, K., Durand, F., Torralba, A.: Learning to predict where humans look. In: 2009 IEEE 12th International Conference on Computer Vision, pp. 2106–2113. IEEE (2009)
29. Kacur, J., Polec, J., Smolejova, E., Heretik, A.: An analysis of eye-tracking features and modelling methods for free-viewed standard stimulus: application for schizophrenia detection. IEEE J. Biomed. Health Inform. **24**(11), 3055–3065 (2020)
30. Krejtz, K., Duchowski, A.T., Niedzielska, A., Biele, C., Krejtz, I.: Eye tracking cognitive load using pupil diameter and microsaccades with fixed gaze. PLoS ONE **13**(9), e0203629 (2018)
31. Kübler, T.C., Rothe, C., Schiefer, U., Rosenstiel, W., Kasneci, E.: Subsmatch 2.0: scanpath comparison and classification based on subsequence frequencies. Behav. Res. Methods **49**, 1048–1064 (2017)
32. Kumar, A., Howlader, P., Garcia, R., Weiskopf, D., Mueller, K.: Challenges in interpretability of neural networks for eye movement data. In: ACM Symposium on Eye Tracking Research and Applications, pp. 1–5 (2020)
33. Locher, P., Krupinski, E., Schaefer, A.: Art and authenticity: behavioral and eye-movement analyses. Psychol. Aesthet. Creat. Arts **9**(4), 356 (2015)
34. Manning, D., Ethell, S., Donovan, T., Crawford, T.: How do radiologists do it? the influence of experience and training on searching for chest nodules. Radiography **12**(2), 134–142 (2006)
35. Massaro, D., et al.: When art moves the eyes: a behavioral and eye-tracking study. PLoS ONE **7**(5), e37285 (2012)

36. Nation, K., Penny, S.: Sensitivity to eye gaze in autism: is it normal? is it automatic? is it social? Dev. Psychopathol. **20**(1), 79–97 (2008)
37. Rajashekar, U., Cormack, L.K., Bovik, A.C., van der Linde, I.: Doves: a database of visual eye movements. Spat. Vis. **22**(2), 161–177 (2009)
38. Sims, S.D., Conati, C.: A neural architecture for detecting user confusion in eye-tracking data. In: Proceedings of the 2020 International Conference on Multimodal Interaction, pp. 15–23 (2020)
39. Tao, Y., Shyu, M.L.: Sp-asdnet: Cnn-lstm based asd classification model using observer scanpaths. In: 2019 IEEE International Conference on Multimedia & Expo Workshops (ICMEW), pp. 641–646. IEEE (2019)
40. Venuprasad, P., et al.: Analyzing gaze behavior using object detection and unsupervised clustering. In: ACM Symposium on Eye Tracking Research and Applications, pp. 1–9 (2020)
41. Vicente, F., Huang, Z., Xiong, X., De la Torre, F., Zhang, W., Levi, D.: Driver gaze tracking and eyes off the road detection system. IEEE Trans. Intell. Transp. Syst. **16**(4), 2014–2027 (2015)
42. Volkmar, F.R., Mayes, L.C.: Gaze behavior in autism. Dev. Psychopathol. **2**(1), 61–69 (1990)
43. Vortmann, L.M., Knychalla, J., Annerer-Walcher, S., Benedek, M., Putze, F.: Imaging time series of eye tracking data to classify attentional states. Front. Neurosci. **15**, 664490 (2021)
44. Wang, Y., Huang, R., Guo, L.: Eye gaze pattern analysis for fatigue detection based on GP-BCNN with ESM. Pattern Recogn. Lett. **123**, 61–74 (2019)
45. Wang, Z., Oates, T., et al.: Encoding time series as images for visual inspection and classification using tiled convolutional neural networks. In: Workshops at the Twenty-ninth AAAI Conference on Artificial Intelligence, vol. 1. AAAI Menlo Park, CA, USA (2015)
46. Zangemeister, W.H., Sherman, K., Stark, L.: Evidence for a global scanpath strategy in viewing abstract compared with realistic images. Neuropsychologia **33**(8), 1009–1025 (1995)
47. Zhang, A.T., Le Meur, B.O.: How old do you look? inferring your age from your gaze. In: 2018 25th IEEE International Conference on Image Processing (ICIP), pp. 2660–2664. IEEE (2018)

Cascading Global and Sequential Temporal Representations with Local Context Modeling for EEG-Based Emotion Recognition

Hyunwook Kang[1], Jin Woo Choi[3], and Byung Hyung Kim[1,2(✉)]

[1] Department of Electrical and Computer Engineering, Inha University, Incheon, Republic of Korea
bhyung@inha.ac.kr
[2] Department of Artificial Intelligence, Inha University, Incheon, Republic of Korea
[3] Department of Neurology and Neurological Sciences, Stanford University School of Medicine, Stanford, CA 94304, USA

Abstract. Electroencephalogram (EEG)-based emotion recognition is an emerging research area in brain-computer interface (BCI) providing a direct window into one's cognitive states. Recent studies employ deep learning models such as a convolutional neural network (CNN), a long short-term memory (LSTM), and the Transformer owing to their high performances achieved for EEG-based emotion recognition. Despite their significant research outcomes, individual networks have their respective limitations in their modeling capabilities. To learn complementary feature representations, we cascade global and sequential temporal representations with local context modeling by unifying CNN, Transformer and LSTM into one framework. To verify the effectiveness of our proposed model, we conducted extensive comparative experiments on two popular benchmark datasets for EEG-based emotion recognition, i.e., SEED-IV, and DEAP, in which we bring further improvements over the recent state-of-the-art models. Our code is publicly available at: https://github.com/affctivai/ConTL.

Keywords: EEG · Emotion Recognition · CNN · Transformer · LSTM

1 Introduction

Emotion recognition is a thriving field of study in artificial intelligence [1,2] due to its strong effects on human cognition [3]. While earlier study in this field has

This work was supported in part by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (RS-2023-00229074, RS-2022-00155915), in part by the National Research Foundation of Korea (NRF) grant funded by the Korea government (No. 2021R1C1C2012437), and in part by INHA UNIVERSITY Research Grant.

A. Antonacopoulos et al. (Eds.): ICPR 2024, LNCS 15313, pp. 305–320, 2025.
https://doi.org/10.1007/978-3-031-78201-5_20

mostly focused on text, vision, and speech [1,4,5], individuals may sometimes conceal their real emotions. On the other hand, physiological signals are more reliable sources as they originate from the central nervous system responding to given stimuli. Among various physiological signals, EEG signals receive the most popular attention since the revelation of correlations between EEG signals and human emotion [6].

Motivated by the effective generalization capability of the convolutional neural network (CNN) [7], many CNN-based deep learning models have been employed for various EEG classification problems [8–10]. Furthermore, a variant of CNN that utilizes the graph neural network called DGCNN [11] has been developed to learn functional connectivity between different EEG channels. Despite their success with significant performances for EEG-based emotion recognition, CNNs are limited in contextual learning due to the fixed size of the receptive field, where the kernel can behold [5,12]. This limitation may not effectively capture important clues for emotion recognition as the human emotional cognitive process continuously evolves [13].

Although a long short-term memory (LSTM) network allows contextual learning by extracting effective temporal relationships using its memory architecture, temporal dependencies are lost after long time steps, which hampers it learning global temporal relationships [4]. Since the advent of the Transformer [14], it has allowed learning global temporal relationships with parallelized computation. However, it is difficult to learn sequential information by the Transformer due to the loss of its position [15]. To complement the respective drawbacks of LSTM and Transformer, a recent method proceeds the Transformer with LSTM [4]. While they have shown a significant performance improvement by complementing the Transformer with LSTM's learned sequential representations, the local context modeling of the convolution operation may bring further improvements for EEG-based emotion recognition as it could detect the emotion-related local patterns [5].

More recently, Conformer [12] uses convolution operation to learn local patterns and further learns global temporal dependencies through the following Transformer [14]. Despite its remarkable performance for different EEG paradigms [12], we suggest that adding sequential relationships can bring more benefits for EEG-based emotion recognition. Therefore, we propose a novel deep learning framework, namely, ConTL, which leverages LSTM to enhance CNN-Transformer network with sequential temporal representations. The proposed model first extracts emotion-related local patterns using convolution operation. Next, the Transformer [14] performs self-attention on the extracted features to learn global temporal information. Then we further cascade the learned features to the following LSTM. For LSTM, we use the stacked bi-directional LSTM (sLSTM) [1] to capture the effective sequential relationships not only from the past but also from the future.

In summary, our contributions are three-folds:

– We explore the effects of learning sequential relationships based on the CNN-Transformer hybrid-network through sLSTM [1] for EEG-based emotion recognition.

- We propose a novel hybrid-network composed of CNN module, Transformer [14] module, and LSTM module for EEG-based emotion recognition.
- We conduct the ablation study to verify the efficacy of our method and show the distribution of the predicted emotional vectors using t-SNE visualization for interpretability [17].

The rest of this paper is organized as follows: In Sect. 2, we begin by exploring existing studies on EEG-based emotion recognition and hybrid-networks related to our proposed network. Section 3 presents our proposed model called ConTL, composed of CNN module, Transformer module, and LSTM module. In Sect. 4, we compare the performance of ConTL with seven state-of-the-art models. In Sect. 5, we articulate our contributions with comprehensive summary. Finally, we draw the conclusion in Sect. 6.

2 Related Works

2.1 EEG-Based Emotion Recognition

There are two types of emotion modeling to measure one's emotional state: discrete emotion modeling and the dimensional emotion modeling [17]. The former defines a set of emotions mainly with six basic categories, i.e., happiness, sadness, disgust, anger, fear, and surprise [18]. The latter measures the degree of emotion with valence and arousal, represented in two-dimensional Cartesian coordinates [17]. The valence ranges from unpleasant to pleasant and the activation level ranging from calm to excited is quantified by the arousal. Their respective degree is positioned on the horizontal axis for valence and the vertical axis for arousal.

Traditional methods for EEG-based emotion recognition have mostly adapted machine learning approaches, which are usually divided into two stages: EEG feature extraction and classifier training [13,17]. There are various methods to extract the EEG features, which can be distinguished by time domain, frequency domain, and time-frequency domain. For example, Wang *et al.*. [17] investigated power spectrum, wavelet, and non-linear dynamical features for EEG signal analysis. While most studies for feature extraction focus on single-channel analysis [11], there are a few methods that attempt to calculate the features from multiple channels to investigate the inter-channel relationships [19,20]. For example, Wu *et al.*. [19] constructed critical sub-network to explore emotion-related functional brain connectivity patterns using three topological features (strength, clustering coefficient, and eigenvector centrality). Li *et al.*. [20] has adapted a multiple feature fusion approach, in which the activation patterns and the connection patterns are combined for emotion recognition. Recently, Kim *et al.*. [21] introduced a discriminative SPD feature learning approach based on Riemannian geometry. Their method normalizes the distribution of SPD matrices and learns the Riemannian center for each class, penalizing the distances between each matrix and its corresponding class center.

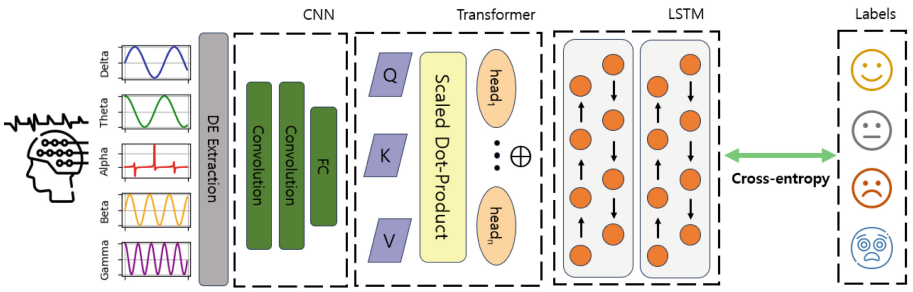


Fig. 1. Illustration of the proposed ConTL's overall architecture.

In our study, we conduct single-channel analysis for EEG-based emotion recognition using deep learning methods. The datasets used in our experiment, SEED-IV [22] and DEAP [23] are based on discrete emotion modeling and dimensional emotion modeling, respectively.

2.2 Hybrid-networks

To date, a hybrid-network has been developed to take advantages of both CNN and LSTM while preventing their respective drawbacks for speech recognition [24]. Sainath *et al.*. [24] proposed CLDNN, which learns temporal structures with reduced frequency variation in the input by placing CNN before LSTM. Shi *et al.*. [25] also proposed a model that combines CNN and LSTM named as ConvLSTM for precipitation nowcasting. Later, their model has been adopted for EEG-based emotion recognition by Kim and Jo [26].

Besides, other previous studies have developed hybrid-networks for EEG-based emotion recognition [12,13,27,28]. Li *et al.*. [27] proposed a hybrid deep learning model called C-RNN that combines inter-channel relationships from CNN with the contextual information from LSTM. Later, a different approach of CNN and LSTM combination has been demonstrated by Yang *et al.*. [28], in which they perform late fusion of the separately learned features from CNN and LSTM. These two different types of CNN and LSTM incorporation are also employed for different EEG paradigms by Zhang *et al.*. [29], in which they defined each method as cascade and parallel methods.

Although there are many other recent deep learning models, they are generally huge with deep layers for good performance, which require large amount of data to prevent overfitting [30]. In contrast, aforementioned hybrid-networks can be used to increase the representational power with small amount of data such as EEG signals due to their difficulty in collection. For a latest model of hybrid-network, Conformer [12] demonstrated superior performance for EEG-based emotion recognition by extracting inter-channel correlations with CNN followed by Transformer for global temporal feature extraction.

3 Methodology

In this section, we first begin by describing the feature extraction method for EEG signals. Then we explain details of our network architecture, which is illustrated in Fig. 1.

3.1 Feature Extraction

Recently, models trained with differential entropy (DE) features have shown superior performances compared to other features for EEG-based emotion recognition [11,31,32]. Thus, we used the extracted DE features from the preprocessed EEG signals as input features to train our classifier [32]. The DE function computes the complexity of continuous random variables as:

$$h(X) = -\int_X f(x_i)\log(f(x_i))dx, \tag{1}$$

where X denotes the set of possible EEG signals and $x_i \in X$ denotes the i-th EEG sample within that range. Instead of directly solving the above (1), we can take the Gaussian distribution $N(\mu, \sigma^2)$ for a random variable to achieve similar outcomes with:

$$\begin{aligned} h(X) &= -\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi\sigma^2}} \exp \frac{(x_i-\mu)^2}{2\sigma^2} \log \frac{1}{\sqrt{2\pi\sigma^2}} \exp \frac{(x_i-\mu)^2}{2\sigma^2} dx \\ &= \frac{1}{2}\log 2\pi e\sigma^2. \end{aligned} \tag{2}$$

In (2), μ, and σ denote the mean and the standard deviation respectively, and we followed the previous method as in [32] to extract the DE features. Firstly, 5 frequency bands (delta: 1–3 Hz, theta: 4–7 Hz, alpha: 8–13 Hz, beta: 14–30 Hz, gamma: 31–50 Hz) are extracted from the EEG signals using a 256-point short-time Fourier transform (STFT). For the window function of STFT, non-overlapping Hann window with 1 s is chosen. Finally, DE features are calculated for each frequency band of every channel.

For feature standardization, we perform the Z-score on the extracted DE features defined as:

$$x_o = \frac{(x_i - \mu)}{\sigma}, \tag{3}$$

where x_o denotes the output of standardization. We note that the mean and the standard deviation are calculated for each frequency band from the train set and their values are directly used to perform the Z-score on the features in the test set.

3.2 Network Architecture

As shown in Fig. 1, ConTL comprises of three different types of network modules, i.e., CNN module, Transformer module, LSTM module, followed by the final classifier for emotion prediction. Each EEG input sample, $x_i \in \mathbb{R}^{C \times F}$, is rearranged into 310-dimensions, where C and F denote the number of channels and the number of frequency bands, respectively. Then it is extended by one dimension for the convolution channel reshaping the EEG input vector into $x_i \in \mathbb{R}^{1 \times 310}$.

CNN Module. To perform local context modeling, the CNN module is composed of two convolutional layers, in which the size of kernel and stride are 4 and 2, respectively. These layers learn the representation of interactions between neighboring electrode channels. For further generalization, the second convolutional layer is followed by LeakyReLU, batch normalization and dropout. The batch normalization drives faster convergence and the dropout is adopted for further regularization. The output channels of these convolutional layers have 64 units and the dropout rate is set to 0.2. Subsequently, these features are downsampled to h units and expanded by one dimension at the first axis to feed them to the following Transformer as tokens.

Transformer Module. The Transformer [14] generates multiple parallel attentions by employing the scaled-dot product defined as:

$$Attention(Q, K, V) = softmax(\frac{QK^T}{\sqrt{h}})V. \tag{4}$$

The Q, K, and V are the tokens from the previous module, which denote the query, key, and value respectively. Given head-specific parameters, $W_i^{q/k/v} \in \mathbb{R}^{h \times h}$, each attention output is computed by leveraging (4) as:

$$head_i = Attention(QW_i^q, KW_i^k, VW_i^v). \tag{5}$$

Then the final output from the Transformer is computed as:

$$\overline{G} = (head_1 \oplus \cdots \oplus head_n)W^o, \tag{6}$$

where $\oplus$ represents the concatenation.

LSTM Module. Once the Transformer outputs the learned global temporal representations, we cascade them to the following LSTM layers as:

$$o_{lstm} = sLSTM(\overline{G}; \theta^{lstm}). \tag{7}$$

The proposed model further learns sequential relationships through these LSTM layers. Each LSTM layer outputs two sequential representations from the

past and the future, where each representation has 8 units. The four sequential temporal representations from two LSTM layers are stacked, whose end-state hidden representations, o_{lstm}, gives 32 units.

Subsequently, one fully connected (FC) layer is used to downsample the output features to M units, in which the M is the number of emotion categories to predict the correct emotion for the given EEG features.

For the loss function, we use the cross-entropy as:

$$\mathcal{L} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{M} y_i \log(\hat{y}_j), \tag{8}$$

where y, and $\hat{y}$ denote the ground-truth and the predicted emotion label, respectively. The symbol N indicates the number of samples in a batch.

To sum up, the convolution operation performs local context modeling by extracting emotion-related local patterns from the given EEG features. These emotion-related local features are cascaded to the following Transformer [14] to learn global temporal relationships. Then the bi-directional LSTM layers complement the output features of Transformer with learning sequential temporal features. Finally, FC layer outputs M units to recognize the corresponding emotion for the given input EEG features.

4 Experiments

In this section, we first describe two datasets used in our experiment, i.e., SEED-IV [22], and DEAP [23]. Next, we delineate the experimental settings. To verify the effectiveness of our proposed ConTL, we compare the mean classification accuracy of all subjects on these two benchmarks with seven baseline models. These baseline models' results are all reproduced by reimplementing them using their public access codes under the same hyper-parameter settings for training. Additionally, we conduct ablation study to observe the effects of learning sequential relationships based on the joint representation of local information and global temporal relationships. We also show the t-SNE visualization of the predicted vectors for interpretability.

4.1 Datasets

SEED-IV [22] contains 45 experiments from 15 participants, in which each participant has taken 3 experiments. The first 16 trials are used for training and the rest 8 trials are used for the test. Each EEG signal is labeled by one of 4 emotion categories, i.e., happy, sad, fear, and neutral.

The DEAP [23] database contains EEG data of 32 participants, which are collected from the 32 electrodes while the participants are watching 1-minute 40 music video clips. Each video has been rated with degrees ranging from 1 to 9 according to the levels of arousal, valence, like / dislike, dominance and familiarity. In this paper, only the EEG signals are used to train the emotion

recognition model without the other 8 peripheral channels. According to the international 10–20 system, 32 channels are chosen to collect the EEG signals. In our experiment, we choose valence and arousal for evaluation.

4.2 Experimental Settings

In our experiments, we prepare consistent train and test sets as in [31] to check whether our reproduced results of DGCNN [11] is reasonable for comparison with our proposed model. While [31] explicitly describes the train set and the test set for the reported mean classification accuracy of DGCNN [11] on SEED-IV [22] dataset, it is unclear whether they have used the same test set for the validation set. However, to prevent data leakage, we used 30% of the original train set for the validation set. As shown in Table 1, our reproduced result of DGCNN [11]'s performance is 65.22%, in which there is a loss of 4.66% compared to what's been reported in [31]. Accounting that this difference may have been resulted from splitting the original train set to apply early stopping guided by the validation loss in our experimental settings, we regarded it can be a good baseline result for comparison with our proposed model.

We have implemented the proposed ConTL with PyTorch and evaluated its performance using the NVIDIA RTX A6000 GPU. To train the model, we used the adam optimizer and set its learning rate, $\beta 1$, $\beta 2$ to 0.0002, 0.5, and 0.999, respectively as demonstrated in [12]. The proposed model is trained for 500 epochs with the batch size of 128. To prevent the overfitting, we use early stopping with an initial patience of 7.

4.3 Baseline Comparison

In this section, we conduct extensive subject-dependent experiments [11,22,31] and compare the performance of the proposed ConTL with seven state-of-the-art methods. For example, EEGNet [8], which is a CNN-based model developed to learn temporal feature perception across different BCI paradigms; CCNN [10] and MT-CNN [9], which have shown remarkable results with CNN-based end-to-end frameworks for EEG-based emotion recognition tasks on the DEAP [23] dataset; DGCNN [11], which exploits CNN with the adjacency matrix to learn functional connectivity between different EEG electrodes for emotion recognition; PCRNN [28], complementing CNN's limited receptive field with learned temporal dependencies from LSTM; Conformer [12], which first mines inter-channel correlations from the input EEG signals with CNN and learns global temporal dependencies through the following Transformer; and DRBN [21], which has improved the understanding of non-stationary EEG signals by learning the barycenters of SPD matrices.

For SEED-IV [22], we use the mean classification accuracy and F1-score of all subjects with their standard deviations to measure the performance. For DEAP [23] in Table 1, we present the mean classification accuracy/standard deviations of valence and arousal, in which their degrees are categorized into low/medium/high states for the classification task.

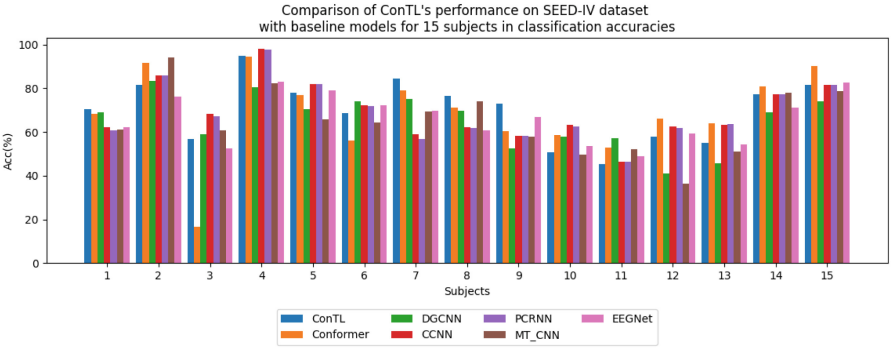


Fig. 2. Comparison of ConTL's performance on the SEED-IV dataset with six baseline models for 15 subjects in classification accuracies.

In Table 1 on SEED-IV [22], the proposed ConTL surpasses the latest model among baselines, DRBN [21], by 1.07% and 1.24% in both mean classification accuracy and F1-score, respectively. Compared to CNN-based end-to-end approach, we can observe that our ConTL significantly improves the classification accuracy by 5.17%, 4.01% over MT-CNN [9] and EEGNet [8]. This result indicates the ConTL's adeptness in capturing cascaded representations of both global and sequential temporal dependencies. Although the mean classification accuracy of EEGNet [8] is not as high as other baseline models but MT-CNN [9], and DGCNN [11], it shows highest mean F1-score over all baseline models. However, our proposed ConTL still outperforms it by 0.93% in mean F1-score, which is an indication that the distribution of our proposed model's predictions are balanced.

Next, we compare classification accuracies for 15 subjects with the six baseline models on SEED-IV [22] as shown in Fig. 2. Compared to Conformer [12],

Table 1. Comparison of the proposed ConTL's performance with baseline models in mean classification accuracies/standard deviations on two benchmarks.

Datasets	SEED-IV		DEAP	
Models	Acc	F1	Valence	Arousal
EEGNet [8]	66.20 / 10.87	69.18 / 11.03	51.45 / 8.59	55.39 / 10.46
CCNN [10]	69.52 / 12.81	68.64 / 14.80	56.22 / 7.15	60.03 / 08.83
MT-CNN [9]	65.04 / 14.36	65.09 / 16.30	54.63 / 7.60	55.86 / 10.01
DGCNN [11]	65.22 / 12.09	65.42 / 12.99	57.81 / 6.73	60.65 / 08.95
PCRNN [28]	69.06 / 13.00	68.27 / 14.98	56.63 / 7.05	60.18 / 08.48
Conformer [12]	68.54 / 18.78	65.83 / 19.16	54.10 / 8.02	57.97 / 10.24
DRBN [21]	69.14 / 11.52	68.87 / 10.36	**58.11** / 5.79	60.93 / 09.38
ConTL	**70.21** / 13.58	**70.11** / 14.27	58.07 / 7.07	**61.06** / 08.82

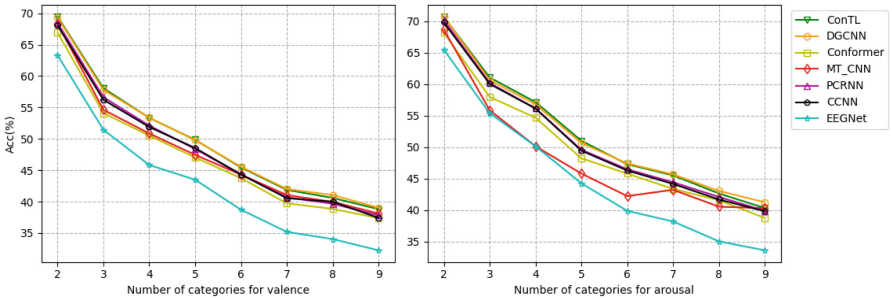


Fig. 3. Comparison of ConTL's performance with six baseline models for eight different categories of valence, arousal on the DEAP dataset in mean classification accuracies.

which learns global temporal dependencies via the Transformer [14] based on CNN, the proposed ConTL yields superior performances over it for 8 number of subjects, i.e., first, third, fourth, fifth, sixth, seventh, eighth, and ninth. While Conformer [12] shows better performances over PCRNN [28] for 10 number of subjects, PCRNN [28] outperforms Conformer [12] for 5 subjects, i.e., the third, fourth, fifth, sixth, and tenth subjects. Although this result is an indication that the Transformer [14] generally performs better than LSTM accounting that Conformer [12] and PCRNN [28] respectively opt the Transformer and LSTM to combine with CNN, there are still some scenarios where LSTM can be more beneficial, in which sequential relationships count. Moreover, for the third subject, Conformer [12] suffers learning discriminative features compared to PCRNN [28]. Although the proposed ConTL's performance for the third subject is not as high as PCRNN [28], we can observe high improvements in ConTL over Conformer [12]. This result underscores the significance of learning both global and sequential temporal information, which can alleviate the performance decrease that might arise by the use of Transformer alone based on CNN. Overall, the proposed ConTL yields superior performances over both PCRNN [28] and Conformer [12] for the first, seventh, eighth, and ninth subjects.

For the DEAP [23] dataset in Table 1, both our innovative ConTL and DRBN [21] outperforms all baseline models for valence and arousal predictions. Compared to DRBN [21], our ConTL shows marginal difference of 0.04% in accuracy for valence prediction, and ConTL surpasses its performance by 0.13% for arousal prediction. DGCNN [11] outperforms all three CNN-based models [8–10] signifying its effectiveness in learning EEG functional connectivity. However, the proposed ConTL outperforms DGCNN [11] by 0.26% and 0.41% for low/medium/high valence and arousal state predictions. PCRNN [28], which jointly fuses learned features from CNN and LSTM, also shows higher performances over all three CNN-based models [8–10] for both valence and arousal predictions on the DEAP dataset. This indicates that CNN can be complemented with additional sequential temporal relationships from LSTM. However, on the SEED-IV dataset [22], PCRNN [28] shows slight reduction in performance

compared to CCNN [10], which indicates that hybrid-networks require different hyper-parameter settings depending on the domains. Similar phenomenon is observed for another type of hybrid-network, Conformer [12], which combines CNN with the Transformer [14]. While Conformer [12] outperforms two CNN-based models, i.e., MT-CNN [9], and EEGNet [8] on SEED-IV [22] by 3.5%, 2.34% in mean classification accuracy, it [12] lags behind MT-CNN [9] by 0.53% for valence prediction and it [12] benefits for arousal prediction with a 2.11% improvement over MT-CNN [9].

To further investigate, we compared the performances of the proposed ConTL with the six baseline models for 8 different categories of degree for valence and arousal from DEAP [23]. We discretized the emotional degree of the valence and arousal by the following rules. Firstly, we calculate the offset by dividing 9 by the chosen number of categories from 2 to 8, whose corresponding offsets are 4.5, 3.0, 2.25, 1.8, 1.5, 1.29, 1.13 respectively. Then the first lower bound starts from 1 and the upper bound is the offset. To compute the next level of the degree, the former upper bound becomes the lower bound and the upper bound is updated by adding the offset. For example, if two categories are chosen for the degree, the first degree level is in between 1 and 4.5 and the next level is in between 4.5 and 9. For three categories of degree, the boundaries of first, second and third levels are in between [1,3.0], [3.0, 6.0], and [6.0, 9.0], respectively. For four categories, the boundaries are as [1, 2.25], [2.25, 4.5], [4.5, 6.75], [6.75, 9.0]. The rest of the categories follow the same rules except for 9 categories. For the final 9 categories, we have rounded up the floating point degree level to the nearest integer.

As expected, similar patterns are shown for different models, in which they suffer from learning discriminative emotional representations as the number of categories increase as shown in Fig. 3. For the arousal prediction in the right of the Fig. 3, it is interesting to observe that the performance of MT-CNN [9] on the prediction of 7 different categories is higher compared to predicting 6 different arousal categories. Compared to the proposed ConTL, we outperform the five baseline models but DGCNN [11]. However, our model generally brings higher improvements over DGCNN [11] for the valence predictions of 2, 3, 5 categories and arousal predictions of 3, 4, 5 categories.

4.4 Ablation Study

In Table 2, we observe the effects of each module in the overall proposed model by removing CNN, Transformer, and LSTM modules, respectively.

On both datasets, significant performance drop is observed by removing CNN module, in which the effects were more significant on SEED-IV over DEAP. While all three modules had their respective effects in the overall performance, the LSTM module contributed to an 0.91% improvement on SEED-IV and 1.24%, 0.85% improvements on DEAP, respectively. This result signifies that the CNN-Transformer network architecture can be improved by learning sequential temporal representations.

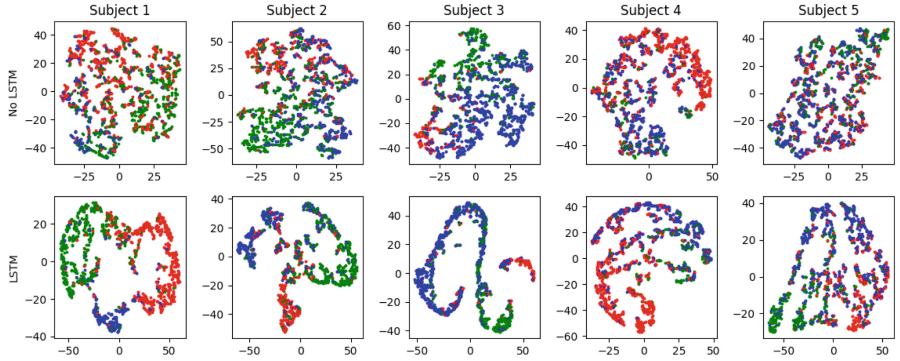


Fig. 4. Visualization of the predicted vectors of valence for the DEAP dataset. The first row shows the distribution of predictions without LSTM and the second row shows the result with LSTM. Different colors represent different emotional categories.

Figure 4 shows the distribution of the predicted vectors embedded into 2-dimensions using t-SNE [17]. While there are many overlappings between different categories without LSTM as shown in the first row of Fig. 4, the proposed model with LSTM helps the predicted vectors to cluster closer to their respective intra-class categories as can be seen in the second row of Fig. 4. Thus, t-SNE [17] visualization further verifies the effectiveness of learning sequential relationships for emotion prediction through LSTM based on the cascaded network of CNN and Transformer.

Next, we change the number of input units to the Transformer, h, which is the number of output units from the CNN module. In Fig. 5, we can observe large perturbations on SEED-IV [22] relative to different values of h. On the other hand, the respective performances relative to different values of h are quite consistent on the DEAP [23] dataset for both valence and arousal compared to SEED-IV [22]. Based on Fig. 5, we chose 50 units for h as it exhibits state-of-the-art performances across two popular benchmark datasets for EEG-based emotion recognition as shown in Table 1.

Table 2. Respective effects of CNN, Transformer, and LSTM modules in the overall model.

Datasets	SEED-IV	DEAP	
Models	Acc	Valence	Arousal
Transformer+LSTM	45.80 / 15.14	56.09 / 7.91	58.91 / 9.02
CNN+LSTM	68.36 / 11.99	57.86 / 6.67	60.35 / 8.36
CNN+Transformer	69.30 / 11.63	56.83 / 7.61	60.21 / 9.13
CNN+Transformer+LSTM	**70.21** / 13.58	**58.07** / 7.07	**61.06** / 8.82

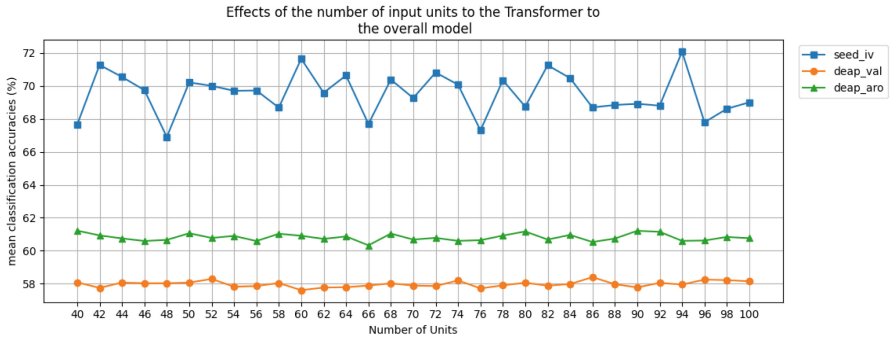


Fig. 5. Effects of the number of input units to Transformer in the overall model.

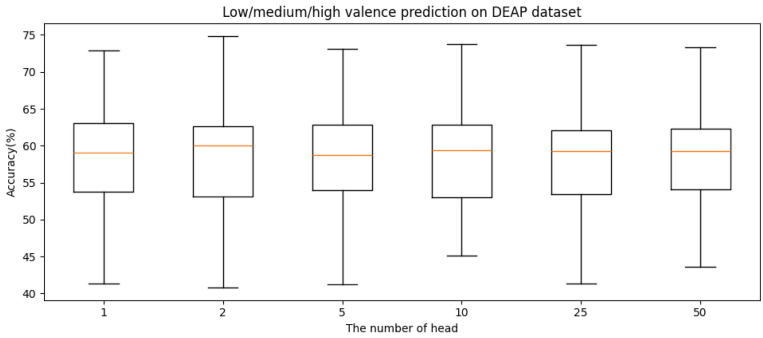


Fig. 6. Effects of the number of attention heads in classification accuracy on DEAP

Different head selections on Transformer networks can learn different aspects of features. Figure 6 presents their relative classification accuracy for low/medium/high valence predictions on DEAP. While we can observe a mild fluctuation in average classification accuracy, they have not significantly influenced in prompting feature learning.

5 Discussion

Previous hybrid-networks for emotion recognition either leverage LSTM or CNN to integrate with the Transformer. While each method lacks local or sequential features, we aim to learn complementary feature representations of global and sequential relationships through Transformer and LSTM based on convolution operation for local context modeling. Through extensive comparative experiments, we verified that the proposed method can learn more discriminative representation over existing CNN-based models, and combined networks of CNN with LSTM or Transformer.

As shown in Fig. 2, we present faithful experimental results by reproducing the baseline models' results using their open access codes. While EEG emotion recognition accuracy on DEAP can be highly improved by using baseline signals [9,28], they can introduce latency for real-time emotion recognition due to the calculation time. Therefore, in our experiments, baseline signals are not used and our comparative experiments' results are only affected by different model architectures but other factors by setting all hyper-parameters same. Although our reproduced results of baseline models could not reach the reported results in [9,28,31], these results are reasonable accounting that we use early stopping with cross-validation. That is, our comparative experiments' results present generalized measure of each model's learning power, in which the performance is only affected by their respective network architectures.

6 Conclusion

This paper presents a novel end-to-end method to learn complementary feature representations by cascading emotion-related local features, global temporal dependencies, and sequential information. The CNN module first learns inter-channel correlations using the convolution operation. The following Transformer [14] performs self-attention on the extracted features to learn global temporal relationships. Subsequently, we leverage sLSTM [1] to further learn sequential relationships from the output of the Transformer [14]. The comparative experiments on two benchmarks show the effectiveness of the proposed ConTL, which surpasses the performance of the state-of-the-art models. To verify the efficacy of learning sequential temporal relationships in addition to local patterns and global temporal dependencies, we conducted the ablation study as shown in Table 2. To further validate the effectiveness of our proposed method, we compared the distribution of the predicted vectors of the proposed ConTL with and without the LSTMs using the t-SNE [17] visualization. While our proposed ConTL yields good performance for emotion recognition from EEG signals, there are still some limitations: 1) the performance of hybrid-networks are dependent on different hyper-parameter settings for different domains; and 2) the large number of parameters induced due to the integration of different types of networks are inevitable, which arise slow inference time during the model deployment. Thus, the following works of hybrid-networks shall consider developing light-weight models robust on different domains. There are many previous studies such as knowledge distillation and domain adaptation, in which they aim to solve these problems and we leave it as our future work.

References

1. Hazarika, D., Zimmermann, R., Poria, S.: Misa: modality-invariant and-specific representations for multimodal sentiment analysis. In: Proceedings of the 28th ACM International Conference on Multimedia, pp. 1122–1131. Association for Computing Machinery, New York, United States (2020)

2. Singh, G.V., Firdaus, M., Chauhan, D.S., Ekbal, A., Bhattacharyya, P.: Zero-shot multitask intent and emotion prediction from multimodal data: a benchmark study. Neurocomputing **569**(127128) (2024)
3. Damasio, A.R.: Descartes' Error: Emotion, Reason, and The Human Brain, 1st edn. Avon Books, New York (1995)
4. Andayani, F., Theng, L.B., Tsun, M.T., Chua, C.: Hybrid LSTM-transformer model for emotion recognition from speech audio files. IEEE Access **10**, 36018–36027 (2022)
5. Zhao, Z., et al.: Combining a parallel 2D CNN with a self-attention Dilated Residual Network for CTC-based discrete speech emotion recognition. Neural Netw. **141**, 52–60 (2021)
6. Zheng, W.L., Zhu, J.Y., Lu, B.L.: Identifying stable patterns over time for emotion recognition from EEG. IEEE Trans. Affect. Comput. **10**(3), 417–429 (2017)
7. Krizhevsky, A., Sutskever, I., Hinton, G. E.: Imagenet classification with deep convolutional neural networks. In: Pereira, F., Burges, C.J., Bottou, L., Weinberger, K.Q. (eds.) Advances in Neural Information Processing Systems, vol. 25 (2012)
8. Lawhern, V.J., Solon, A.J., Waytowich, N.R., Gordon, S.M., Hung, C.P., Lance, B.J.: EEGNet: a compact convolutional neural network for EEG-based brain-computer interfaces. J. Neural Eng. **15**(5), 056013 (2018)
9. Rudakov, E., et al.: Multi-Task CNN model for emotion recognition from EEG Brain maps. In: 4th International Conference on Bio-Engineering for Smart Technologies, pp. 1–4. IEEE, Paris, France (2021)
10. Yang, Y., Wu, Q., Fu, Y., Chen, X.: Continuous convolutional neural network with 3D input for EEG-based emotion recognition. In: Cheng, L., Leung, A., Ozawa, S. (eds.) Neural Information Processing. ICONIP 2018, LNCS, vol. 11307, pp. 433–443. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-04239-4_39
11. Song, T., Zheng, W., Song, P., Cui, Z.: EEG emotion recognition using dynamical graph convolutional neural networks. IEEE Trans. Affect. Comput. **11**(3), 532–541 (2020)
12. Song, Y., Zheng, Q., Liu, B., Gao, X.: EEG conformer: convolutional transformer for EEG decoding and visualization. IEEE Trans. Neural Syst. Rehabil. Eng. **31**, 710–719 (2022)
13. Li, X., et al.: EEG based emotion recognition: a tutorial and review. ACM Comput. Surv. **55**(4), 1–57 (2022)
14. Vaswani, A., et al.: Attention is all you need. Adv. Neural Inf. Process. Syst. **30** (2017)
15. Dai, Z., Yang, Z., Yang, Y., Carbonell, J., Le, Q., Salakhutdinov, R.: Transformer-XL: attentive language models beyond a fixed-length context. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pp. 2978–2988. Association for Computational Linguistics, Florence (2019)
16. Van der Maaten, L., Hinton, G.: Visualizing data using t-SNE. J. Mach. Learn. Res. **9**(11) (2008)
17. Wang, X.W., Nie, D., Lu, B.L.: Emotional state classification from EEG data using machine learning approach. Neurocomputing **129**, 94–106 (2014)
18. Ekman, P.: An argument for basic emotions. Cogn. Emot. **6**(3–4), 169–200 (1992)
19. Wu, X., Zheng, W.L., Lu, B.L.: Identifying functional brain connectivity patterns for EEG-based emotion recognition. In: 9th International IEEE/EMBS Conference on Neural Engineering, pp. 235–238. IEEE, San Francisco, USA (2019)
20. Li, P., Liu, H., Si, Y., Li, C., Li, F., Zhu, X., et al.: EEG based emotion recognition by combining functional connectivity network and local activations. IEEE Trans. Biomed. Eng. **66**(10), 2869–2881 (2019)

21. Kim, B.H., Choi, J.W., Lee, H., Jo, S.: A discriminative SPD feature learning approach on Riemannian manifolds for EEG classification. Pattern Recognit. **143** (2023)
22. Zheng, W.L., Liu, W., Lu, Y., Lu, B.L., Cichocki, A.: Emotionmeter: a multimodal framework for recognizing human emotions. IEEE Trans. Cybern. **49**(3), 1110–1122 (2018)
23. Koelstra, S., Muhl, C., Soleymani, M., Lee, J.S., Yazdani, A., Ebrahimi, T., et al.: DEAP: a database for emotion analysis; using physiological signals. IEEE Trans. Affect. Comput. **3**(1), 18–31 (2011)
24. Sainath, T.N., Vinyals, O., Senior, A., Sak, H.: Convolutional, long short-term memory, fully connected deep neural networks. In: 2015 IEEE International Conference on Acoustics. Speech and Signal Processing, pp. 4580–4584. IEEE, South Brisbane, Australia (2015)
25. Shi, X., Chen, Z., Wang, H., Yeung, D.Y., Wong, W.K., Woo, W.C.: Convolutional LSTM network: a machine learning approach for precipitation nowcasting. In: Cortes, C., Lawrence, N., Lee, D., Sugiyama, M., Garnett, R. (eds.) Advances in Neural Information Processing Systems, vol. 28. Curran Associates Inc., Montreal, Canada (2015)
26. Kim, B.H., Jo, S.: Deep physiological affect network for the recognition of human emotions. IEEE Trans. Affect. Comput. **11**(2), 230–243 (2018)
27. Li, X., Song, D., Zhang, P., Yu, G., Hou, Y., Hu, B.: Emotion recognition from multi-channel EEG data through convolutional recurrent neural network. In: 2016 IEEE International Conference on Bioinformatics and Biomedicine, pp. 352–359. IEEE, Shenzhen, China (2016)
28. Yang, Y., Wu, Q., Qiu, M., Wang, Y., Chen, X.: Emotion recognition from multi-channel EEG through parallel convolutional recurrent neural network. In: 2018 International Joint Conference on Neural Networks, pp. 1–7. IEEE, Rio de Janeiro, Brazil (2018)
29. Zhang, D., et al.: Cascade and parallel convolutional recurrent neural networks on EEG-based intention recognition for brain computer interface. In: Williams, B., Chen, Y., Neville, J. (eds.) Proceedings of the AAAI Conference on Artificial Intelligence, vol. 32, AAAI Press, Washington, DC, USA (2018). https://doi.org/10.1609/aaai.v32i1.11496
30. Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., Salakhutdinov, R.: Dropout: a simple way to prevent neural networks from overfitting. J. Mach. Learn. Res. **15**(1), 1929–1958 (2014)
31. Zhong, P., Wang, D., Miao, C.: EEG-based emotion recognition using regularized graph neural networks. IEEE Trans. Affect. Comput. **13**(3), 1290–1301 (2022)
32. Zheng, W.L., Lu, B.L.: Investigating critical frequency bands and channels for EEG-based emotion recognition with deep neural networks. IEEE Trans. Auton. Ment. Dev. **7**(3), 162–175 (2015)

Engagement Measurement Based on Facial Landmarks and Spatial-Temporal Graph Convolutional Networks

Ali Abedi[1(✉)] and Shehroz S. Khan[1,2]

[1] KITE Research Institute, University Health Network, Toronto, Canada
{ali.abedi,shehroz.khan}@uhn.ca
[2] Institute of Biomedical Engineering, University of Toronto, Toronto, Canada

Abstract. Engagement in virtual learning is crucial for a variety of factors including student satisfaction, performance, and compliance with learning programs, but measuring it is a challenging task. There is therefore considerable interest in utilizing artificial intelligence and affective computing to measure engagement in natural settings as well as on a large scale. This paper introduces a novel, privacy-preserving method for engagement measurement from videos. It uses facial landmarks, which carry no personally identifiable information, extracted from videos via the MediaPipe deep learning solution. The extracted facial landmarks are fed to Spatial-Temporal Graph Convolutional Networks (ST-GCNs) to output the engagement level of the student in the video. To integrate the ordinal nature of the engagement variable into the training process, ST-GCNs undergo training in a novel ordinal learning framework based on transfer learning. Experimental results on two video student engagement measurement datasets show the superiority of the proposed method compared to previous methods with improved state-of-the-art on the EngageNet dataset with a 3.1% improvement in four-class engagement level classification accuracy and on the Online Student Engagement dataset with a 1.5% improvement in binary engagement classification accuracy. Gradient-weighted Class Activation Mapping (Grad-CAM) was applied to the developed ST-GCNs to interpret the engagement measurements obtained by the proposed method in both the spatial and temporal domains. The relatively lightweight and fast ST-GCN and its integration with the real-time MediaPipe make the proposed approach capable of being deployed on virtual learning platforms and measuring engagement in real-time.

Keywords: Engagement Measurement · Graph Convolutional Network · Ordinal Classification · Transfer Learning

1 Introduction

Engagement is key in education, blending students' attention and interest within a learning context [1]. It not only stems from existing interests but also fosters

A. Antonacopoulos et al. (Eds.): ICPR 2024, LNCS 15313, pp. 321–338, 2025.
https://doi.org/10.1007/978-3-031-78201-5_21

new ones through sustained attention [2], essential for content comprehension and engagement development [1]. However, measuring and upholding engagement is challenging, and demands significant effort from educators. The advent of remote sensing, Artificial Intelligence (AI), and affective computing offers new avenues for accurately measuring engagement across various learning environments, including virtual learning platforms. Technologies such as facial expression recognition [3] and eye gaze tracking [4] enable more precise monitoring and enhancement of student engagement. This paper explores the development of AI algorithms for automated engagement measurement in virtual learning, taking into account the definition of engagement in educational psychology and its measurement methodologies.

Fredricks et al. [5] defined engagement through affective, behavioral, and cognitive components. Affective engagement relates to students' emotions and attitudes towards their tasks, including interest or feelings during a learning class [6]. Behavioral engagement refers to active participation, such as focusing on a computer screen or not being distracted by a phone [7]. Cognitive engagement deals with a student's dedication to learning and tackling challenges, affecting outcomes such as recall and understanding [1]. Booth et al. [1] identified that engagement is measured through various signals such as facial expressions, eye movements, posture, heart rate, brain activity, audio, and digital interactions. Cameras, prevalent in devices for online learning, make video a key data modality for AI-based engagement measurement [8,9]. Therefore, AI techniques predominantly focus on video to measure student engagement [8–10].

AI-driven engagement measurement methods are divided into end-to-end and feature-based approaches. While feature-based methods generally outperform end-to-end approaches, which process raw videos, they require extensive identification of effective features through trial and error [8–10]. Extracting multi-modal features from videos with multiple computer vision algorithms or neural networks [11,12], and their subsequent analysis by deep neural networks to measure engagement, render these methods computationally demanding [8,13]. Such computational requirements limit their use on local devices and necessitate the transfer of privacy-sensitive video data to cloud servers for engagement measurement. In contrast, extracting low-dimensional landmark information, such as facial and hand landmarks, not only provides more compact data but also captures essential geometric features for affect and behavior analysis, including engagement, without personal identifiers [14–16]. Previous feature-based engagement measurement approaches extracted features such as head pose, facial Action Units (AUs), iris and gaze features, and affect and emotion features, which are indicators of the behavioral and affective components of engagement [8–10]. The literature has demonstrated the success of facial landmarks in capturing these aforementioned features [3,17–20]. A model such as Spatial-Temporal Graph Convolutional Networks (ST-GCNs) [21], capable of analyzing spatial-temporal facial landmarks and learning these aforementioned features inherently, can reduce the need for raw facial videos. An approach based on facial landmarks prioritizes pri-

vacy and reduces computational demands, making it more practical for real-time engagement measurement.

In this paper, an alternative course away from conventional end-to-end and feature-based approaches is charted, and a novel engagement measurement technique using facial landmarks extracted from videos is presented. The proposed method is characterized by its privacy-preserving nature and computational efficiency. This work makes the following contributions:

- This marks the first instance in video-based engagement measurement [8,9,13,22–25] where facial landmarks, as the single data modality extracted from videos, are analyzed through ST-GCNs [21] to infer the engagement level in the video;
- To integrate the ordinal nature of the engagement variable into the training process, ST-GCNs undergo training in a novel ordinal learning framework utilizing transfer learning;
- Extensive experiments conducted on two video-based engagement measurement datasets demonstrate the superiority of the proposed method over previous methods, achieving an improved state-of-the-art in engagement level classification accuracy. For explainability, Gradient-weighted Class Activation Mapping (Grad-CAM) is applied to the developed ST-GCNs to interpret the engagement measurements obtained by the proposed method in both spatial and temporal domains.

2 Related Work

The literature review in this paper serves two main purposes: discussing past video-based engagement measurement techniques and examining the use of graph convolutional networks for facial expression and affect analysis.

2.1 Engagement Measurement

Karimah et al. [8] conducted a systematic review on measuring student engagement in virtual learning environments, revealing a focus on affective and behavioral components of engagement. Most methods utilize datasets annotated by external observers [10] to train both feature-based and end-to-end models. In the following, some of the relevant feature-based and end-to-end works on video-based engagement measurement are discussed.

In the domain of end-to-end engagement measurement techniques, deep neural networks analyze consecutive raw video frames to output the engagement level of the student in the video. These methods do not employ the extraction of handcrafted features from the videos; instead, the network is adept at autonomously learning to extract the most useful features directly from the videos, utilizing consecutive convolutional layers. The deep neural networks implemented in these end-to-end approaches include networks capable of video analysis, such as 3D Convolutional Neural Networks (3D CNNs) and Video

Transformers [26–28], as well as combinations of 2D CNNs with sequential neural networks such as Long Short-Term Memory (LSTM) and Temporal Convolutional Network (TCN) [26,27,29,30].

The process of measuring engagement through feature-based techniques involves two stages. Initially, behavioral and affective features are extracted from video frames, relying on either domain-specific knowledge or pre-trained models for facial embedding extraction. OpenFace [31] is notably effective for its comprehensive feature extraction capabilities, including AUs, eye movements, gaze direction, and head positioning, and is widely applied in engagement measurement [12,13,32,33]. Examples of facial embedding include the Masked Autoencoder for facial video Representation LearnINg (MARLIN) [11], utilized by Singh et al. [12], and the Emotion Face Alignment Network (EmoFAN) [19], used by Abedi et al. [13]. Subsequently, to analyze these extracted features and infer engagement, various machine learning and deep learning models are employed, such as Bag-of-Words (BoW) [32], Recurrent Neural Network (RNN) variations [24], Temporal Convolutional Networks (TCNs) [13,33], Transformers [12,34], and ensemble models [35].

2.2 Graph-Based Facial Affect and Expression Analysis

Liu et al. [3] conducted a comprehensive review of the literature on graph-based methods for facial affect analysis. These methods typically take an image or a sequence of images as input and produce an affect classification or regression as output. Based on their review, Liu et al. [3] proposed a pipeline for graph-based facial affect analysis, which includes (i) face preprocessing, (ii) graph-based affective representation, and (iii) graph-based relational reasoning. Preprocessing involves steps such as face detection and registration. Graph-based affective representation involves defining the structure of the graph, i.e., nodes and edges. The graph structure can be spatial or spatiotemporal depending on whether the input data is still images or videos. The graph structure can be at the landmark level, region level, or AU level, with nodes representing facial landmarks, facial regions of interest, or facial AUs, respectively. In the relational reasoning step, the edges, nodes, their interrelations, and temporal dynamics are analyzed through relational reasoning machine-learning or deep-learning models to make inferences regarding affect. Models used to analyze graph data include dynamic Bayesian networks, RNNs, CNNs, fully-connected neural networks, and non-temporal and temporal graph neural networks [3].

Zhou et al. [36] proposed a facial expression recognition method based on spatiotemporal landmark-level and region-level graphs. The intra-frame graph is formed by the connections among thirty-four facial landmarks located around the eyes, lips, and cheeks. The definition of these intra-frame connections was done manually. Inter-frame connections were established by linking each node in one frame to its corresponding node in the following frame. Two parallel ST-GCNs with analogous structures were trained; one on the nodes' x- and y-coordinates, and another on their histogram of orientation features. ST-GCNs'

outputs were concatenated and processed by a fully connected network to identify facial expressions. This method's disadvantages include independent training of the two ST-GCNs rather than joint learning, and manual definition of nodes and edges.

Wei et al. [15] presented a graph-based method for micro-expression recognition in video. The method included a dual-stream ST-GCN, focusing on the x and y coordinates of facial landmarks and the distance and angles between adjacent facial landmarks. An AU-specific loss function was incorporated into the neural network's training process in order to incorporate the association between AUs and micro-expressions. Their methodology employed three different sets of facial landmarks as graph nodes, comprising sets with 14, 31, and Dlib's [37] 68 facial landmarks. Notably, the set with 68 landmarks was the only one to include landmarks along the jawline. The experiments demonstrated the best results for micro-expression recognition when 14 facial landmark sets were used.

Leong et al. [25] introduced a method employing spatial-temporal graph attention networks to analyze facial landmarks and head poses, aimed at identifying academic emotions. The facial landmarks used were those around the eyebrow, eye, nose, and mouth excluding those that outline the face's outer shape, with the reasoning being they lack correlation with affective states. A notable limitation of this approach is its reliance on multiple deep neural networks for extracting features, i.e., facial landmarks and head pose. The method achieved lower academic emotion detection accuracy when compared to prior feature-based methods [13].

2.3 Discussion

As reviewed in this section, while some methods have used facial landmarks and ST-GCNs to recognize facial affect and expression in videos, there has been no exploration of the use of these methods to measure engagement. The fundamental differences between engagement and facial affect and expression make their measurement different. First, engagement is a multi-component state comprising behavioral, affective, and cognitive components. To illustrate, key indicators of behavioral engagement, such as being off-task or on-task, are determined by head pose, eye gaze, and blink rate. These indicators, and therefore engagement, cannot be effectively measured using methods designed solely for facial affect analysis. Second, engagement is not a constant state; it varies over time and should be measured at specific time resolutions where it remains stable and quantifiable. An ideal measurement duration for engagement is between ten and forty seconds, which is longer than the time resolution for facial affect analysis, which sometimes occurs at the frame level. Third, engagement measurement could involve recognizing an ordinal variable that indicates levels of engagement, as opposed to facial expression recognition, which identifies categorical variables without inherent order.

Existing engagement measurement approaches face limitations due to the necessity of employing multiple deep neural networks for the extraction and

analysis of multi-modal features. Coupled with the significant differences between facial affect analysis and engagement measurement as outlined above, these limitations underscore a gap in the field. In response, we introduce a straightforward yet effective method for engagement measurement. This method is lightweight and fast, preserving privacy while also demonstrating improvements over current methodologies across two video-based engagement measurement datasets.

3 Method

The input to the proposed method is a video sample of a student seated in front of a laptop or PC camera during a virtual learning session. The sequences of facial landmarks extracted from consecutive video frames through MediaPipe [14,20] are analyzed by ST-GCN [16,21] to output the engagement level of the student in the video.

3.1 Graph-Based Representation

The MediaPipe deep learning solution [14], a framework that is both real-time and cross-platform, is employed for the extraction of facial landmarks from video. Incorporated within MediaPipe, Attention Mesh [20] is adept at detecting 468 3D facial landmarks throughout the face and an extra 10 landmarks for the iris. However, not all 478 landmarks are employed in the proposed engagement measurement method. Consistent with existing studies [3], only 68 of the 3D facial landmarks, which match those identified by the Dlib framework [37], in addition to the 10 3D iris landmarks, making a total of 78 landmarks, are utilized.

The 3D facial landmarks encapsulate crucial spatial-temporal information pertinent to head pose [17], AUs [18], eye gaze [20], and affect [3,19]-key features identified in prior engagement measurement studies [8–10,13,22–24,26,27,33]. Consequently, there is no necessity for extracting additional handcrafted features from the video frames.

The N 3D facial landmarks extracted from T consecutive video frames are utilized to construct a spatiotemporal graph $G = (V, E)$. In this graph, the set of nodes $V = \{v_{ti} | t = 1, \ldots, T, i = 1, \ldots, N\}$ encompasses all the facial landmarks in a sequence. To construct G, first, the facial landmarks within one frame are connected with edges according to a connectivity structure based on Delaunay triangulation [38] which is consistent with true facial muscle distribution and uniform for different subjects [3]. Then each landmark will be connected to the same landmark in the consecutive frame.

3.2 Graph-Based Reasoning

Based on the spatiotemporal graph of facial landmarks $G = (V, E)$ constructed above, an adjacency matrix A is defined as an $N \times N$ matrix where the element at position (i, j) is set to 1 if there is an edge connecting the i^{th} and j^{th} landmarks, and set to 0 otherwise. An identity matrix I, of the same dimensions as A,

is created to represent self-connections. A spatial-temporal graph, as the basic element of an ST-GCN layer, is implemented as follows [21].

$$f_{\text{out}} = \left(\Lambda^{-\frac{1}{2}}((A+I)\odot M)\Lambda^{-\frac{1}{2}} f_{\text{in}} W_{\text{spatial}}\right) W_{\text{temporal}} \tag{1}$$

where $\Lambda^{ii} = \sum_j (A^{ij} + I^{ij})$. M is a learnable weight matrix that enables scaling the contributions of a node's feature to its neighboring nodes [21]. The input feature map, denoted f_{in}, is the raw coordinates of facial landmarks for the first layer of ST-GCN, and it represents the outputs from previous layers in subsequent layers of ST-GCN. The dimensionality of f_{in} is (C, N, T), where C is the number of channels, for example, 3 in the initial input to the network corresponding to the x, y, and z coordinates of the facial landmarks. In each layer of ST-GCN, initially, the spatial (intra-frame) convolution is applied to f_{in} based on the weight matrix W_{spatial}, utilizing a standard 2D convolution with a kernel size of 1×1. Subsequently, the resulting tensor is multiplied by the normalized adjacency matrix $\Lambda^{-\frac{1}{2}}((A+I)\odot M)\Lambda^{-\frac{1}{2}}$ across the spatial dimension. Afterward, the temporal (inter-frame) convolution, based on the weight matrix W_{temporal}, is applied to the tensor output from the spatial convolution. This convolution is a standard 2D convolution with a kernel size of $1 \times \Gamma$, where Γ signifies the temporal kernel size.

Following an adequate number of ST-GCN layers, specifically three in this work, that perform spatial-temporal graph convolutions as outlined above, the resulting tensor undergoes 2D average pooling. The final output of the network is generated by a final 2D convolution. This convolution employs a kernel size of 1×1 and features an output channel dimensionality equal to the number of classes K, i.e., the number of engagement levels to be measured. An explanation of the detailed architecture of the ST-GCNs for specific datasets can be found in Subsect. 4.2.

3.3 Ordinal Engagement Classification Through Transfer Learning

The model described above, with a final layer comprising K output channels, tackles the engagement measurement problem as a categorical K-class classification problem without taking into account the ordinal nature of the engagement variable [10,39,40]. The model could harness the ordinality of the engagement variable to enhance its inferences. Drawing inspiration from [13,41], a novel ordinal learning framework based on transfer learning is introduced as follows.

Training phase- The original K-level ordinal labels, $y = 0, 1, \ldots, K-1$, in the training set are converted into $K-1$ binary labels y_i as follows: if $y > i$, then $y_i = 1$; otherwise, $y_i = 0$, for $i = 0, 1, \ldots, K-2$. Subsequently, $K-1$ binary classifiers are trained with the training set and the $K-1$ binary label sets described above. The training of binary classifiers is based on transfer learning, which proceeds as follows: Initially, a network is trained on the dataset with the original K-class labels, employing a regular final layer with K output channels. After training, the ST-GCN layers of this network are frozen, and the final layer is removed. To this frozen network, $K-1$ separate untrained 2D convolution layers

with a single output channel each are added, resulting in $K-1$ new networks. Each of these networks consists of a frozen sequence of ST-GCN layers followed by an untrained 2D convolution layer. These $K-1$ new networks are then trained on the entire dataset using the $K-1$ binary label sets described above. During this phase, only the final 2D convolution layers are subjected to training.

Inference phase- For the ordinal classification of a test sample, the sample is initially input into the pre-trained (and frozen) sequence of ST-GCN layers, followed by a 2D average pooling layer. The tensor obtained from this process is then input into $K-1$ pre-trained final 2D convolution layers, each yielding a probability estimate for the test sample being in the binary class $y_t = y_i$, where $i = 0, 1, \ldots, K-2$. Subsequently, these $K-1$ binary probability estimates are transformed into a single multi-class probability of the sample belonging to class $y = 0, 1, \ldots, K-1$, as follows [41].

$$p(y_t = k) = \begin{cases} 1 - p(y_t \geq 0), & \text{if } k = 0, \\ p(y_t > k-1) - p(y_t \geq k), & \text{if } 0 < k < K-1, \\ p(y_t > K-2), & \text{if } k = K-1. \end{cases} \tag{2}$$

Despite the increased training time for the ordinal model within the aforementioned ordinal learning framework, the final count of parameters in the ordinal model remains nearly identical to that of the original non-ordinal model.

4 Experiments

This section evaluates the performance of the proposed method relative to existing methods in video-based engagement measurement. It reports and discusses the results of multi-class and binary classification of engagement across two datasets. Based on the engagement measurement problem at hand, several evaluation metrics are employed. In the context of multi-class engagement level classification, metrics such as accuracy and confusion matrix are reported. For binary engagement classification, accuracy, the Area Under the Curve of the Receiver Operating Characteristic (AUC-ROC), and the Area Under the Curve of the Precision and Recall curve (AUC-PR) are utilized. In addition, the number of parameters of the models used, memory consumption, and inference time of the proposed method are compared to those of previous methods.

4.1 Datasets

Experiments on two large video-based engagement measurement datasets were conducted, each presenting unique challenges that further enabled the validation of the proposed method.

EngageNet: The EngageNet dataset [12], recognized as the largest dataset for student engagement measurement, includes video recordings of 127 subjects participating in virtual learning sessions. Each video sample has a duration of 10 s, with a frame rate of 30 fps and a resolution of 1280×720 pixels. The

subjects' video recordings were annotated as four ordinal levels of engagement: Not-Engaged, Barely-Engaged, Engaged, and Highly-Engaged. The dataset was divided into 7983, 1071, and 2257 samples for training, validation, and testing, respectively, using a subject-independent data split approach [12]. However, only the training and validation sets were made available by the dataset creators and were utilized for the training and validation of predictive models in the experiments presented in this paper. The distribution of samples in the four aforementioned classes of engagement in the training and validation sets are 1550, 1035, 1658, and 3740, and 132, 97, 273, and 569, respectively.

Online SE: The Online SE dataset [42], comprises videos of six students participating in online courses via the Zoom platform. These recordings span 10 s each, with a frame rate of 24 fps and a resolution of 220×155 pixels. The videos were annotated as either Not-Engaged or Engaged. The dataset was segmented into 3190, 1660, and 1290 samples for training, validation, and testing, respectively. The distribution of samples in the two aforementioned classes of engagement in the training, validation, and test sets is 570 and 2620, 580 and 1080, and 570 and 720, respectively.

4.2 Experimental Setting

The sole information extracted from video frames is facial landmarks, which are analyzed by ST-GCN to determine the engagement level of the student in the video. Drawing inspiration from the pioneering works on body-joints-based action analysis [21,43], the proposed ST-GCN for facial-landmarks-based engagement measurement is structured as follows. The input facial landmarks are first processed through a batch normalization layer, followed by three consecutive ST-GCN layers with 64, 128, and 256 output channels, respectively. Residual connections are incorporated in the last two ST-GCN layers. A dropout rate of 0.1 is applied to each ST-GCN layer. The temporal kernel size in the ST-GCN layers is selected to be 9. Subsequent to the ST-GCN layers, an average pooling layer is utilized, and its resulting tensor is directed into a 2D convolutional layer with 256 input channels and a number of output channels corresponding to the number of classes. A Softmax activation function then computes the probability estimates. In cases where engagement measurement is framed as a binary classification task, the terminal 2D convolution layer is configured with a single output channel, substituting Softmax with a Sigmoid function. The Sigmoid function is also employed for the individual binary classifiers within the ordinal learning framework detailed in Subsect. 3.3. The models are trained using the Adam optimizer with mini-batches of size 16 and an initial learning rate of 0.001 for 300 epochs. The learning rate is decayed by a factor of 0.1 every 100 epochs.

4.3 Experimental Results

Comparison to Previous Methods. Table 1 presents the comparative results of two settings of the proposed method, a regular non-ordinal classifier and an ordinal classifier, with previous methods on the validation set of the EngageNet

Table 1. Classification accuracy of engagement levels on the validation set of the EngageNet dataset [12]: comparison of state-of-the-art end-to-end methods and feature-based methods with various feature sets and classification models against two configurations of the proposed method - facial landmarks analyzed by ST-GCN and ordinal ST-GCN. Bolded values denote the best results.

Ref.	Features	Model	Accuracy
[27]	End to End Model	ResNet + TCN	0.5472
[30]	End to End Model	EfficientNet + LSTM	0.5757
[30]	End to End Model	EfficientNet + Bi-LSTM	0.5894
[12]	Gaze	LSTM	0.6125
[12]	Head Pose	LSTM	0.6760
[12]	AU	LSTM	0.6303
[12]	Gaze + Head Pose	LSTM	0.6769
[12]	Gaze + Head Pose + AU	LSTM	0.6704
[12]	Gaze	CNN-LSTM	0.6060
[12]	Head Pose	CNN-LSTM	0.6732
[12]	AU	CNN-LSTM	0.6172
[12]	Gaze + AU	CNN-LSTM	0.6275
[12]	Head Pose + AU	CNN-LSTM	0.6751
[12]	Gaze + Head Pose + AU	CNN-LSTM	0.6751
[12]	Gaze	TCN	0.6256
[12]	Head Pose	TCN	0.6611
[12]	AU	TCN	0.6293
[12]	Gaze + Head Pose + AU	TCN	0.6779
[12]	Gaze	Transformer	0.5545
[12]	Gaze + Head Pose	Transformer	0.6445
[12]	Gaze + Head Pose + AU	Transformer	0.6910
[12]	Gaze + Head Pose + AU + MARLIN	Transformer	0.6849
[34]	Gaze	TCCT-Net	0.6433
[34]	Head Pose	TCCT-Net	0.6891
[34]	AU	TCCT-Net	0.6629
[34]	Gaze + Head Pose	TCCT-Net	0.6564
[34]	Gaze + Head Pose + AU	TCCT-Net	0.6713
Ours	Facial Landmarks	ST-GCN	0.6937
Ours	Facial Landmarks	Ordinal ST-GCN	**0.7124**

dataset [12]. The engagement measurement in EngageNet [12] is a four-class classification problem and the accuracy is reported as the evaluation metric. The previous methods in Table 1 include state-of-the-art end-to-end methods, including the combination of ResNet-50 with TCN [27] and the combination of EfficientNet B7 with LSTM and bidirectional LSTM [30], followed by state-of-the-art feature-based methods. The previous feature-based methods in Table 1

used different combinations of OpenFace's eye gaze, head pose, and AU features [31] along with MARLIN's facial embedding features [11]. These features were classified by LSTM, CNN-LSTM, TCN, and Transformer. Refer to [12] for more details. The results of the feature-based method proposed by Vedernikov et al. [34] are also reported, where the aforementioned features are classified using a Transformer-based neural network called the Tensor-Convolution and Convolution-Transformer Network (TCCT-Net).

Despite the abundant data samples in the EngageNet dataset [12] available for training complex neural networks such as ResNet + TCN [27] and EfficientNet B7 + bidirectional LSTM [30], their performance is inferior to that of feature-based methods. This highlights the necessity of extracting hand-crafted features or facial landmarks from videos and building classifiers on top of them.

In the single feature configurations of previous methods in Table 1, head pose achieves better results compared to AUs, which is better than eye gaze. While head pose and eye gaze are indicators of behavioral engagement [7,13], AUs, which are associated with facial expressions and affect [19], are indicators of affective engagement. Combining these three features is always beneficial since engagement is a multi-component variable that can be measured by affective and behavioral indicators when the only available data is video [13]. Among classification models, utilizing more advanced models improves accuracy; the Transformer is better than TCN, which is better than CNN-LSTM and LSTM; however, it comes at the cost of increased computational complexity. For single features, the Transfomer-based TCCT-Net [34] outperforms other classifiers; however, for multiple feature sets, the vanilla Transformer [12] outperforms the others.

The proposed ST-GCN in Table 1, which relies solely on facial landmarks without requiring raw facial videos or multiple hand-crafted features, outperforms previous methods. Moreover, making the proposed method ordinal further improves the state-of-the-art by 3.1% compared to eye gaze, head pose, AUs, and MARLIN features [11] with the Transformer [12]. Tables 2a and 2b depict the confusion matrices of the ordinal and non-ordinal configurations of the proposed method in Table 1. As shown, incorporating ordinality significantly increases the number of correctly classified samples in the first three classes and results in a 2.7% improvement in accuracy compared to its non-ordinal counterpart.

Table 2. Confusion matrices of the proposed method with (a) non-ordinal and (b) ordinal ST-GCN on the validation set of the EngageNet dataset [12].

(a) Non-ordinal ST-GCN

Class	**1**	**2**	**3**	**4**
1	99	14	13	6
2	10	29	33	25
3	9	19	100	145
4	5	4	45	515

(b) Ordinal ST-GCN

Class	**1**	**2**	**3**	**4**
1	104	12	10	6
2	15	36	24	22
3	10	26	112	125
4	9	5	44	511

Table 3. Classification accuracy of engagement levels on the validation set of the EngageNet dataset [12] for different variants of the proposed method.

Variant	Accuracy
ST-GCN is replaced with LSTM	0.6847
ST-GCN is replaced with TCN	0.6813
Only x and y coordinates of joints are used	0.6655
A temporal kernel of 3 is used	0.6748
A temporal kernel of 15 is used	0.6841
Every 2 frames are used	0.6813
Every 4 frames are used	0.6907
Every 8 frames are used	0.6907
Every 16 frames are used	0.6841
Hand landmarks are added	0.6956

The feature extraction step in [12,34] involves running multiple deep-learning models in OpenFace [31] to capture eye gaze, head pose, and AUs, as well as another complex network for MARLIN feature embeddings [11]. While these feature extraction networks are complex for real-time use, the proposed method relies on facial landmarks extracted using the real-time MediaPipe [14,20]. Considering only the classification models, the number of parameters in EfficientNet B7 + LSTM [30], ResNet + TCN [27], Transformer [12], the proposed non-ordinal ST-GCN, and ordinal ST-GCN are 82,681,812, 24,639,236, 1,063,108, 861,688, and 861,431, respectively. The memory consumption for EfficientNet B7 + LSTM [30], ResNet + TCN [27], Transformer [12], and the proposed ordinal ST-GCN are 2268.78, 790.35, 178.00, and 180.96 megabytes, respectively. The inference time for classifying a data sample using EfficientNet B7 + LSTM [30], ResNet + TCN [27], Transformer [12], and the proposed ordinal ST-GCN are 348, 51, 10, and 0.8 milliseconds, respectively. This indicates the efficiency of the proposed method, which, while being lightweight and fast, also improves the state-of-the-art.

Variants of the Proposed Method. Table 3 displays the results of different variants of the proposed method on the validation set of the EngageNet [12] dataset. In the first two variants, the x, y, and z coordinates of facial landmarks are converted into multivariate time series and analyzed by an LSTM and TCN. The LSTM includes four unidirectional layers with 256 neurons in hidden units and is followed by a 256×4 fully connected layer. The parameters of the TCN are as follows: the number of layers, number of filters, kernel size, and dropout rate are 8, 64, 8, and 0.05, respectively. While their results are acceptable and better than most of the earlier methods in Table 1, they cannot outperform ST-GCN, the last two rows of Table 1. The fact that the accuracy of the LSTM and TCN in Table 3 is higher than those in Table 1 signifies the efficiency of

facial landmarks for engagement measurement. In the third variant, the z coordinates of facial landmarks are disregarded, and the decrease in the accuracy of the non-ordinal ST-GCN indicates the importance of the z coordinates for engagement measurement. Temporal kernel sizes other than 9 in the fourth and fifth variants have a negative impact on the results of the non-ordinal ST-GCN. When engagement measurement is performed using every 2, 4, 8, and 16 frames, instead of every frame, there is a slight decrease in the accuracy of the non-ordinal ST-GCN. However, this is a trade-off between accuracy and computation since reducing the frame rate corresponds to a reduction in computation. The last row of Table 3 shows the results of the ordinal ST-GCN when 21 hand landmarks extracted using MediaPipe [14] were added to the facial landmarks. The lower accuracy compared to using only facial landmarks is due to two factors. Firstly, there is 80% missingness in the hand landmarks due to the absence of hands in the videos. Secondly, it indicates that facial landmarks alone are sufficient for engagement measurement, capturing both behavioral and emotional indicators of engagement.

Results on the Online SE Dataset. Table 4 presents the results of the proposed method in comparison to previous methods on the test set of the Online SE dataset [42]. The earlier methods listed in Table 4 are feature-based, extracting affective and behavioral features from video frames and performing binary classification of the features using TCN in [33], TCN in [13], LSTM with attention in [44], and BoW in [32]. Given that engagement measurement in the Online SE dataset [42] is posed as a binary classification problem, implementing the ordinal version of the proposed method is not required. The proposed method, employing facial landmarks with ST-GCN, attains the highest accuracy and AUC PR.

Table 4. Binary classification accuracy, AUC ROC, and AUC PR of engagement on the test set of the Online SE dataset [42]: comparison of previous methods against the proposed method. Bolded values denote the best results.

Method	Accuracy	AUC ROC	AUC PR
[33]	0.7803	0.8764	0.8008
[13]	0.7637	0.8710	0.7980
[44]	0.7475	0.8890	0.7973
[32]	0.8191	**0.8926**	0.9018
Proposed	**0.8315**	0.8806	**0.9131**

Interpretation of Results. Figure 1 displays the interpretation of engagement measurements taken using the proposed method by applying Grad-CAM [43] to the ordinal ST-GCN trained on the training set of the EngageNet dataset. For visualization purposes, three exemplary frames (out of 300) from the beginning,

middle, and end of three data samples annotated as Not-Engaged in the validation set of the EngageNet dataset are shown in Fig. 1 (a)-(c). The facial landmarks extracted through MediaPipe are overlaid on the frames, where the color map of the facial landmarks, from blue to red, depicts the class activation map values of the last ST-GCN layer in the trained ST-GCN. The facial landmarks associated with the target class of Not-Engaged at certain frames are colored towards red. In Fig. 1 (a), during the beginning and middle of the video, the first two exemplary frames show the student engaged, with lower class activation map values. At the end of the video, when the student is not looking at the camera (computer screen) and is looking elsewhere, landmarks on the iris, eye, and jawline show higher values, resulting in the model classifying the sample as Not-Engaged. In Fig. 1 (b), the student is not paying attention and is playing with their phone. Jawline, eye, iris, and eyebrow facial landmarks with red colors have higher class activation map values, resulting in the model classifying the sample as Not-Engaged. In Fig. 1 (c), throughout the entire video sample, the head pose of the student is normal and perpendicular to the camera. However, the eyes are almost closed, indicating sleepiness and low arousal. This is detected by the higher class activation map values on the eye and iris landmarks, corresponding to the intensity of AU number 45, which indicates how closed the eyes are. This resulted in the model classifying the sample as Not-Engaged.

5 Discussion

Our research led to the development of a novel deep-learning framework for student engagement measurement. In the proposed framework, sequences of facial landmarks are extracted from consecutive video frames and analyzed by ordinal ST-GCNs to make inferences regarding the engagement level of the student in the video. The successful application of our model to the EngageNet [12] and Online SE [42] datasets not only confirms its efficacy but also establishes a new standard in engagement level classification accuracy, outperforming previous methods. As the sole input information to the developed engagement measurement models, the 3D facial landmarks contain information about head pose [17], AUs [18], eye gaze [20], and affect [3,19], which are the key indicators of behavioral and affective engagement. The relatively lightweight and fast ST-GCN and its integration with real-time MediaPipe [14] make the proposed framework capable of being deployed on virtual learning platforms and measuring engagement in real-time. The proposed method is privacy-preserving and does not require access to personally identifiable raw video data for engagement measurement. In a real-world deployment, the cross-platform MediaPipe solution [14], running on a web, mobile, or desktop application, extracts facial landmarks from video data on users' local devices. These non-identifiable facial landmarks are then transferred to a cloud, where they are analyzed by ST-GCNs to measure engagement. The interpretability feature of the proposed method, enabled through Grad-CAM, facilitates understanding which facial landmarks, corresponding to behavioral and affective indicators of engagement, contribute to certain levels of engagement. It also helps identify the specific timestamps at which these contributions

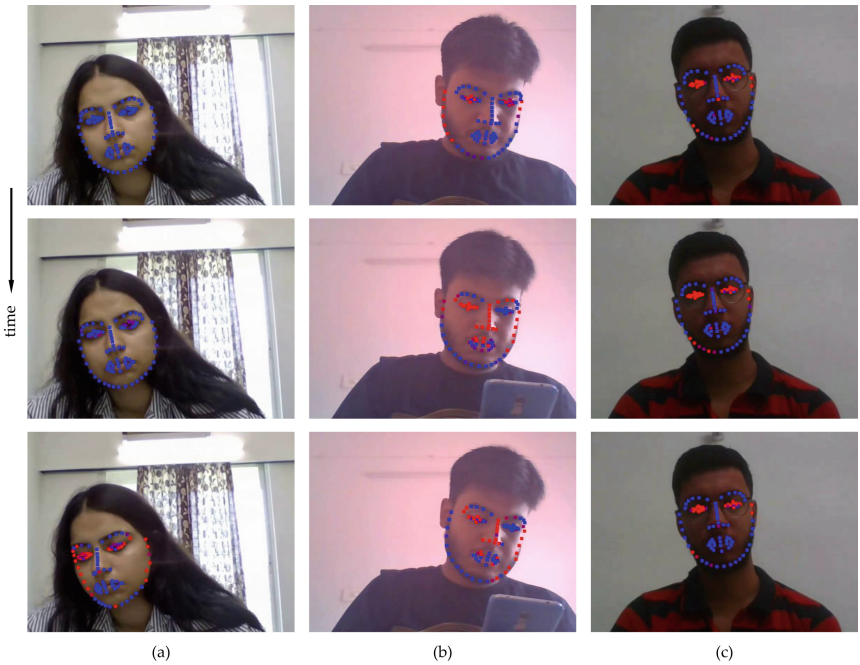


Fig. 1. Interpretation of engagement measurements taken using the proposed method by applying Gradient-weighted Class Activation Mapping (Grad-CAM) to the Spatial-Temporal Graph Convolutional Network (ST-GCN). Please refer to the last paragraph of Subsect. 4.3 for further details.

occur. This provides instructors with additional information to take necessary actions and promote student engagement. A limitation of the proposed method is its reliance on the quality of facial landmarks detected by MediaPipe. In the context of engagement measurement in virtual learning sessions, an occluded or absent face, and consequently non-detected facial landmarks, correspond to lower levels of engagement or disengagement. Our developed ST-GCN was able to correctly classify most samples with occluded or absent faces, i.e., no facial landmarks, as Not-Engaged. To improve the performance of the proposed method, the following direction could be investigated: analyzing facial landmarks with more advanced ST-GCNs, which are equipped with attention mechanisms and trained through contrastive learning techniques and applying augmentation techniques to video data before facial landmark extraction [45] or to facial landmark data to improve the generalizability of ST-GCNs.

Acknowledgment. The authors express their sincere thanks to the Multimodal Perception Lab at the International Institute of Information Technology, Bangalore, India, for their generosity in providing the Online SE dataset, which was instrumental in the execution of our experiments.

This research was funded by the Natural Sciences and Engineering Research Council of Canada.

References

1. Booth, B.M., Bosch, N., D'Mello, S.K.: Engagement detection and its applications in learning: a tutorial and selective review. Proc. IEEE **111**(10), 1398–1422 (2023). https://doi.org/10.1109/JPROC.2023.3309560
2. Hidi, S., Renninger, K.A.: The four-phase model of interest development. Educ. Psychol. **41**(2), 111–127 (2006)
3. Y. Liu, X. Zhang, Y. Li, J. Zhou, X. Li, and G. Zhao: Graph-based facial affect analysis: a review. IEEE Trans. Affect. Comput. (2022)
4. Wood, E., Baltrusaitis, T., Zhang, X., Sugano, Y., Robinson, P., Bulling, A.: Rendering of eyes for eye-shape registration and gaze estimation. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 3756–3764 (2015)
5. Fredricks, J.A., Blumenfeld, P.C., Paris, A.H.: School engagement: potential of the concept, state of the evidence. Rev. Educ. Res. **74**(1), 59–109 (2004)
6. S. D'Mello and A. Graesser: dynamics of affective states during complex learning.Learning and Instruction **22**(2), 145–157 (2012)
7. J. Ocumpaugh: Baker Rodrigo Ocumpaugh monitoring protocol (BROMP) 2.0 technical and training manual. New York, NY and Manila, Philippines: Teachers College, Columbia University and Ateneo Laboratory for the Learning Sciences 60 (2015)
8. Karimah, S.N., Hasegawa, S.: Automatic engagement estimation in smart education/learning settings: a systematic review of engagement definitions, datasets, and methods. Smart Learn. Environ. **9**(1), 1–48 (2022)
9. M. Dewan, M. Murshed, and F. Lin: Engagement detection in online learning: a review. Smart Learn. Environ. **6**(1), 1–20 (2019)
10. S.S. Khan, A. Abedi, and T. Colella: Inconsistencies in Measuring Student Engagement in Virtual Learning-A Critical Review. arXiv preprint arXiv:2208.04548 (2022)
11. Cai, Z., et al.: Marlin: masked autoencoder for facial video representation learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 1493–1504 (2023)
12. Singh, M., Hoque, X., Zeng, D., Wang, Y., Ikeda, K., Dhall, A.: Do i have your attention: a large scale engagement prediction dataset and baselines. In: Proceedings of the 25th International Conference on Multimodal Interaction. ICMI 2023, pp. 174–182. Association for Computing Machinery, New York, NY, USA (2023)
13. Abedi, A., Khan, S.S.: Affect-driven ordinal engagement measurement from video. Multimedia Tools Appl. (2023)
14. Lugaresi, C., et al.: Mediapipe: a framework for building perception pipelines. arXiv preprint arXiv:1906.08172 (2019)
15. Wei, J., Peng, W., Lu, G., Li, Y., Yan, J., Zhao, G.: Geometric graph representation with learnable graph structure and adaptive au constraint for micro-expression recognition. IEEE Trans. Affect. Comput. (2023)
16. Zheng, K., Wu, J., Zhang, J., Guo, C.: A skeleton-based rehabilitation exercise assessment system with rotation invariance. IEEE Trans. Neural Syst. Rehabil. Eng. **31**, 2612–2621 (2023)
17. Malek, S., Rossi, S.: Head pose estimation using facial-landmarks classification for children rehabilitation games. Pattern Recogn. Lett. **152**, 406–412 (2021)
18. Jacob, G.M., Stenger, B.: Facial action unit detection with transformers. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 7680–7689 (2021)

19. Toisoul, A., Kossaifi, J., Bulat, A., Tzimiropoulos, G., Pantic, M.: Estimation of continuous valence and arousal levels from faces in naturalistic conditions. Nat. Mach. Intell. **3**(1), 42–50 (2021)
20. Grishchenko, I., Ablavatski, A., Kartynnik, Y., Raveendran, K., Grundmann, M.: Attention mesh: high-fidelity face mesh prediction in real-time. arXiv preprint arXiv:2006.10962 (2020)
21. 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 (2018)
22. Ma, X., Xu, M., Dong, Y., Sun, Z.: Automatic student engagement in online learning environment based on neural turing machine. Int. J. Inf. Educ. Technol. **11**(3), 107–111 (2021)
23. Copur, O., Nakip, M., Scardapane, S., Slowack, J.: Engagement detection with multi-task training in e-learning environments. In: International Conference on Image Analysis and Processing, pp. 411–422 (2022)
24. Abedi, A., Khan, S.S.: Detecting disengagement in virtual learning as an anomaly using temporal convolutional network autoencoder. Sig. Image Video Process. (2023)
25. Fwa, H.L.: Fine-grained detection of academic emotions with spatial temporal graph attention networks using facial landmarks (2022)
26. Gupta, A., D'Cunha, A., Awasthi, K., Balasubramanian, V.: Daisee: towards user engagement recognition in the wild. arXiv preprint arXiv:1609.01885 (2016)
27. Abedi, A., Khan, S.S.: Improving state-of-the-art in detecting student engagement with Resnet and TCN hybrid network. In: 2021 18th Conference on Robots and Vision (CRV), pp. 151–157 (2021)
28. Ai, X., Sheng, V.S., Li, C.: Class-attention video transformer for engagement intensity prediction. arXiv preprint arXiv:2208.07216 (2022)
29. Liao, J., Liang, Y., Pan, J.: Deep facial spatiotemporal network for engagement prediction in online learning. Appl. Intell. **51**(10), 6609–6621 (2021)
30. Selim, T., Elkabani, I., Abdou, M.A.: Students engagement level detection in online e-learning using hybrid efficientnetb7 together With TCN, LSTM, and Bi-LSTM. IEEE Access **10**, 99573–99583 (2022)
31. Baltrusaitis, T., Zadeh, A., Lim, Y.C., Morency, L.-P.: Openface 2.0: facial behavior analysis toolkit. In: 2018 13th IEEE International Conference on Automatic Face & Gesture recognition (FG 2018), pp. 59–66 (2018)
32. Abedi, A., Thomas, C., Jayagopi, D.B., Khan, S.S.: Bag of states: a non-sequential approach to video-based engagement measurement. arXiv preprint arXiv:2301.06730 (2023)
33. Thomas, C., Nair, N., Jayagopi, D.B.: Predicting engagement intensity in the wild using temporal convolutional network. In: Proceedings of the 20th ACM International Conference on Multimodal Interaction, pp. 604–610 (2018)
34. Vedernikov, A., Kumar, P., Chen, H., Seppänen, T., Li, X.: TCCT-Net: TwoStream network architecture for fast and efficient engagement estimation via behavioral feature signals. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 4723–4732 (2024)
35. Tian, X., Nunes, B.P., Liu, Y., Manrique, R.: Predicting student engagement using sequential ensemble model. IEEE Trans. Learn. Technol. (2023)
36. Zhou, J., Zhang, X., Liu, Y., Lan, X.: Facial expression recognition using spatial-temporal semantic graph network. In: 2020 IEEE International Conference on Image Processing (ICIP), pp. 1961–1965 (2020). https://doi.org/10.1109/ICIP40778.2020.9191181

37. MISC
38. Liu, Y., Zhang, X., Lin, Y., Wang, H.: Facial expression recognition via deep action units graph network based on psychological mechanism. IEEE Trans. Cogn. Dev. Syst. **12**(2), 311–322 (2019)
39. Yannakakis, G.N., Cowie, R., Busso, C.: The ordinal nature of emotions: an emerging approach. IEEE Trans. Affect. Comput. **12**(1), 16–35 (2018)
40. Whitehill, J., Serpell, Z., Lin, Y.-C., Foster, A., Movellan, J.R.: The faces of engagement: automatic recognition of student engagementfrom facial expressions. IEEE Trans. Affect. Comput. **5**(1), 86–98 (2014)
41. Frank, E., Hall, M.: A simple approach to ordinal classification. In: Machine Learning: ECML 2001: 12th European Conference on Machine Learning Freiburg, Germany, September 5-7, 2001 Proceedings 12, pp. 145–156 (2001)
42. Thomas, C., Sarma, K.P., Gajula, S.S, Jayagopi, D.B.: Automatic prediction of presentation style and student engagement from videos. Comput. Educ. Artif. Intell. (2022)
43. Zheng, K., Wu, J., Zhang, J., Guo, C.: A skeleton-based rehabilitation exercise assessment system with rotation invariance. IEEE Trans. Neural Syst. Rehabil. Eng. (2023)
44. Chen, X., Niu, L., Veeraraghavan, A., Sabharwal, A.: FaceEngage: robust estimation of gameplay engagement from user-contributed (YouTube) videos. IEEE Trans. Affect. Comput. (2019)
45. Abedi, A., Malmirian, M., Khan, S.S.: Cross-modal video to body-joints augmentation for rehabilitation exercise quality assessment. arXiv preprint arXiv:2306.09546 (2023)

A Spatial-Temporal Graph Convolutional Network for Video-Based Group Emotion Recognition

Xingzhi Wang[1,2], Tao Chen[1], and Dong Zhang[1(✉)]

[1] Sun Yat-sen University, Guangzhou 510006, China
zhangd@mail.sysu.edu.cn
[2] Guangdong University of Finance and Economics, Guangzhou 510320, China

Abstract. There are complex emotional interactions between individuals in group and between group and individuals. Although existing methods for group emotion recognition (GER) made quite efforts to learn the spatial-temporal characteristics via efficient deep learning-based networks, they neglected to model the interactive characteristics within group videos. In this article, we propose a graph-based GER approach to learn the spatial-temporal interactive characteristics from the facial and holistic cues of a group video. Specifically, we construct a spatial-temporal graph with facial information to describe the emotional relationships within a group in spatial and temporal dimensions. We employ a graph attention network (GAT) to dynamically model the emotional relationships and influences between individual and group nodes across the spatial-temporal dimension. The proposed method utilizes the GAT to explore the temporal correlations of holistic features extracted from the video frames. The introduced graph attention mechanism helps the proposed network effectively focus on the important nodes, capture interactive information, and generate a more precise spatial-temporal representation for GER. We fuse the decisions based on facial and holistic information in a linear way to obtain a comprehensive recognition result for the emotional state of group videos. Extensive experiments demonstrate that the proposed method learns effective spatial-temporal emotional features, and achieves superior performance in overall accuracies of 70.23% and 92.90% on the VGAF and GECV datasets, respectively.

Keywords: Spatial-temporal Graph · Group Emotion Recognition · Graph Attention Network

1 Introduction

Group emotion plays a crucial role in human society, influencing individual performance in areas such as cooperation, conflict resolution, creativity, and social cohesion [1]. Accurate identification of group emotions is beneficial for promoting cooperation and achieving better outcomes. In the field of affective computing, most research on group emotion recognition (GER) focuses on estimating

A. Antonacopoulos et al. (Eds.): ICPR 2024, LNCS 15313, pp. 339–354, 2025.
https://doi.org/10.1007/978-3-031-78201-5_22

group emotional states from static images of individuals in multiple social events. Although existing image-based GER approaches have achieved promising results [4,19,20], they are incompetent to learn the emotional cues implied in successive frames of group video.

Recently, video-based GER methods were proposed to estimate the group emotional state during a period time. Current video-based GER methods have proved the feasibility and effectiveness of estimating group emotions from videos. However, the emotional interaction existed within the group has not been taken into consideration in these method. Researches on social psychology have demonstrated that there are various emotional interaction relationships within a group, such as the emotional influence between group members, the individual's impact on the generation of the group emotion, and the group emotion feedback on individuals [8]. Moreover, this phenomenon of emotional interaction not only exists at each moment within the group, but also manifests over time. For instance, the group's emotion at a moment affects the subsequent emotional states of individuals and the group. This fact requires an efficient GER method employed to describe the spatial emotional relationships of a group at each moment and characterize the temporal relationship of emotion in multiple consecutive instantaneous frames.

In this work, we propose a graph-based approach to learn the interactive emotional cues from a group video for video-based GER. We construct a graph data structure to describe the complex emotional relationships among group videos. Specifically, this work represents each group member at any given moment as a node (individual node), and the instantaneous group emotion at each moment as a pseudo node (group node). These directional edges in the graph indicate the emotional influence relationships between nodes at the same moment (*i.e.*, spatial relationship) or the emotional influence relationships between nodes over time (*i.e.*, temporal relationship). Furthermore, for effective usage of the constructed graphs of group videos, we employ a graph attention network (GAT) to dynamically learn the interaction information between different nodes, capture the interaction patterns between nodes, and generate effective spatial-temporal representations for group emotion recognition. Overall, the main contributions of our paper can be summarized as:

- We construct spatial-temporal relational graph data and comprehensively describe the complex emotional relationships over a specific time period within a group.
- We introduce the attention mechanism to exploit the constructed spatial-temporal graph data and learn the weights of importance between nodes in the graph. The proposed method can prioritize nodes with significant emotional impact and extract interactive information more effectively.
- Extensive experiments on two popular video-based GER benchmarks, Video level Group AFfect (VGAF) [15], and Group-level Emotion on Crowd Videos (GECV) [14] datasets, show that the proposed method outperformed the state-of-the-art methods. The proposed method improved the best accuracies

of GER from 68.02% to 70.23% on the VGAF dataset and from 90.46% to 92.90% on the GECV dataset.

2 Related Work

Video-based group emotion recognition (GER) is a process that involves analyzing and interpreting the collective emotional state of a group of individuals in a video. Existing video-based GER methods employ temporal models to characterize the correlations between consecutive features across frames, and yield temporal representations of group emotions. For instance, Sharma et al. [10] proposed the VGAFNet net-work to extract facial features from each video frame and generate temporal emotional fea-tures with an LSTM model for group emotional recognition. Liu et al. [10] used a TSM temporal model [9] and an OpenSmile toolkit [3] to extract group emotional features from audio and dynamic sequences, respectively. Additionally, some research utilized 3D convolutional neural networks (3D CNNs) to capture spatial-temporal emotional information from group videos [13]. The existing video-based GER methods have shown impressive recognition performance, while these approaches neglect to model the fact that there exist group member interactions in a group and these interactions influence group emotion. The simulations of emotional interactions rarely are reflected in existing research, and the effectiveness of learned emotional features has room for improvement. In this work, we aim to model the spatial-temporal emotional interaction among the group video and extract interactive information for accurate group emotion recognition.

3 Proposed Approach

The proposed approach constructs two graphs by learned facial and holistic information to represent the relationship implied in a group video. Then, we use a graph attention network (GAT) with the constructed graphs to model the emotional interactions from individual to individual, individual to group, and group to individual temporally. Modeling emotional interactions among the group video helps the model learn interactive information and obtain effective temporal representation for GER. Finally, the GER results obtained from facial and holistic views are linearly integrated as a comprehensive recognition result. The overall framework of our approach is illustrated in Fig. 1.

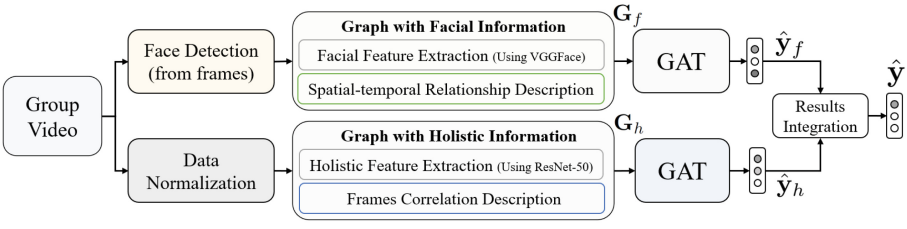


Fig. 1. The overall framework of the proposed approach

3.1 Problem Formulation

Video-based group emotion recognition involves identifying and classifying the collective emotional state of a group of people from video data. Let $\mathbf{V}$ denotes the group video, and $\mathbf{y} = [y_1, y_2, \ldots, y_C]$ denotes the emotional label of the sample, where C is the number of emotion categories. Assuming each group video consists of T frames, and $\mathbf{V} = \left[\mathbf{I}^1, \mathbf{I}^2, \ldots, \mathbf{I}^T\right]$. Consequently, the holistic feature sequence extracted from T frame images can be represented as a feature matrix $\mathbf{X}_h = \left[\mathbf{h}^1, \mathbf{h}^2, \ldots, \mathbf{h}^T\right]$, where $\mathbf{h}$ denotes the holistic feature of a frame image. We use the graph data structure to describe the emotional relationship for each frame of a group video. In the proposed method, group members are regarded as nodes (individual nodes), and the edge connecting each two individual nodes indicates the emotional relationship between the two individuals. The attribute (feature) of the i-th individual node in the t-th frame image is denoted as $\mathbf{x}_i^t$. $\mathbf{X} = \left[\mathbf{x}_1^1, \mathbf{x}_2^1, \ldots, \mathbf{x}_N^T\right]$ represents the feature matrix formed by all individual features in the group video. The emotional relationship between individuals in the t-th frame image can be characterized by the adjacency matrix $\mathbf{A}^t$. In this article, we assume the edge in the graph is fully connected and the weight of an edge is learned from the node's attributes. Therefore, adjacency matrix $\mathbf{A}^t$ is an $N \times N$ matrix of all ones, $\mathbf{1}_{N\times N}$, where N is the number of individuals in the video. For a group video $\mathbf{V}$, there exists a set which contains T adjacency matrices and the set can be denoted as $\mathbf{A}_v = \left\{\mathbf{A}^1, \mathbf{A}^2, \ldots, \mathbf{A}^T\right\}$.

3.2 Feature Extraction

Facial information in a group offers direct insights into the emotional states of its members and is the most significant cue for GER. Besides, holistic information includes humans, surrounding objects, and background, and it also provides rich emotional cues for estimating the emotional state of a group. Considering the effectiveness of facial and holistic information for GER, the proposed approach separately extracts the facial and holistic features from a frame image to represent the individual and global emotional state.

To extract effective individual emotion features, we use the fine-tuned VGGFace network [12], which has been trained on face images and emotion labels, as the facial emotion extraction network. The output of the last fully connected layer of the VGGFace network is used as the emotional feature (a 4096-dimensional vector) for each individual node. The feature extraction process of the i-th individual node in the t-th frame image is expressed by Eq. (1),

$$\mathbf{x}_i^t = \text{VGGFace}\left(\mathbf{I}_i^t\right), \tag{1}$$

where $\mathbf{I}_i^t$ denotes the i-th face image in the t-th frame image, and $\mathbf{x}_i^t \in \mathbb{R}^{4096}$ is the individual emotion feature extracted from the face image. We employ the popular face detector MTCNN [21] to detect and crop the regions of faces from each frame of group videos. The sequence number of the individual in each image frame is determined by sorting the output confidence of the face detector network.

The effectiveness of holistic information has been demonstrated in group emotion recognition tasks [15]. To effectively capture emotional cues from frame images, we employ the ResNet-50 network [5] trained on image-based GER datasets [2,5], to extract the holistic emotional features:

$$\mathbf{h}^t = \text{ResNet-50}\left(\mathbf{I}^t\right), \tag{2}$$

where $\mathbf{I}^t$ denotes the t-th frame image of a group video. We use the output of the last convolutional layer of the ResNet-50 network as the holistic emotional feature (a 2048-dimensional vector), and $\mathbf{h}^t \in \mathbb{R}^{2048}$ is the holistic emotion feature extracted from the t-th frame image.

3.3 Graph with Facial Information

Previous works described the emotional relationship within a group image and benefited from exploiting interactive information from spatial relations for image-based GER [10,18]. However, the emotional cues implied in the temporal relationships have not been explored in video-based GER. In this work, we introduce a spatial-temporal emotional graph to comprehensively characterize the complex relationship in the group video.

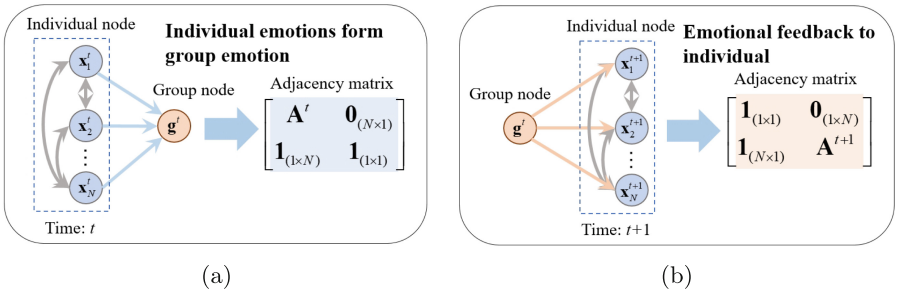


Fig. 2. Emotional interaction considering temporal relationships. (a) The state of the group node at time t generated by the aggregation of individual node. (b) The state of the group node at time t affects the states of individual nodes at time t+1

1) **Basic emotional relationships**: In order to characterize the emotional influence relation in both space and time, we introduce a pseudo (group) node into the emotional relationship graph at every moment. As illustrated in Fig. 2, we aggregate the attribution of individual nodes at the current moment to generate the attribution of the group node (instantaneous group emotion). Then, the instantaneous group emotion serves as an emotional context [8] to influence the attributions of individual nodes at the subsequent moment. Figure 2 illustrates three types of emotional relation: the emotional influence between individuals at each moment; the impact of the individuals on the instantaneous group emotion; and the influence of the generated instantaneous group emotion on individuals

at the next moment. Figure 2(a) depicts an emotional relationship that individual emotions influence the group emotion, and it can be represented by the adjacency matrix as Eq. (3),

$$\mathbf{A}_{x\to g}^{t} = \begin{bmatrix} \mathbf{A}^{t} & \mathbf{0}_{(N\times 1)} \\ \mathbf{1}_{(1\times N)} & \mathbf{1}_{(1\times 1)} \end{bmatrix}, \tag{3}$$

where $\mathbf{A}_{x\to g}^{t}$ denotes the adjacency matrix of the individual node and group node at time t, x denotes the attribution of the individual node, g denotes the attribution of group node, and the symbol $x \to g$ for aggregating the information of individual nodes to generate new attribution of the group node. In Eq. (3), $\mathbf{A}^{t}$ represents the individual emotional relation at time t, $\mathbf{0}_{(N\times 1)}$ denotes an $N \times 1$ dimensional vector consisting entirely of zeros, and $\mathbf{1}_{(1\times N)}$ denotes a $1\times N$ dimensional vector consisting entirely of ones. The adjacency matrix $\mathbf{A}_{x\to g}^{t}$ indicates the unidirectional relationship between N individual nodes and the group node.

Figure 2(b) depicts an emotional relationship where the group emotion influences individuals at the next moment, and it can be represented by the adjacency matrix shown in Eq. (4),

$$\mathbf{A}_{g\to x}^{t} = \begin{bmatrix} \mathbf{1}_{(1\times 1)} & \mathbf{0}_{(1\times N)} \\ \mathbf{1}_{(N\times 1)} & \mathbf{A}^{t+1} \end{bmatrix}, \tag{4}$$

where $\mathbf{A}_{g\to x}^{t}$ denotes the adjacency matrix of a group node at the moment t and an individual node at the moment t+1, the symbol $g \to x$ denotes the group node influences the individual node on emotion. $\mathbf{A}^{t+1}$ represents the individual emotional relation at time t+1.

2) **Spatial-temporal Graph in Serial Pattern**: Eqs. (3) and (4) represent the adjacency matrixes of two types of emotional relationships, which are temporally adjacent in sequence. To build the spatial-temporal emotional relationship, we define the merging operation for the adjacency matrices $\mathbf{A}_{x\to g}^{t}$ and $\mathbf{A}_{g\to x}^{t}$ as shown in Eq. (5),

$$\begin{aligned} \mathbf{A}_{x\to g}^{t} \cup \mathbf{A}_{g\to x}^{t} &= \begin{bmatrix} \mathbf{A}^{t} & \mathbf{0}_{(N\times 1)} & \text{zero padding} \\ \mathbf{1}_{(1\times N)} & \mathbf{1}_{(1\times 1)} & \mathbf{0}_{(1\times N)} \\ \text{zero padding} & \mathbf{1}_{(N\times 1)} & \mathbf{A}^{t+1} \end{bmatrix} \\ &= \begin{bmatrix} \mathbf{A}^{t} & \mathbf{0}_{(N\times 1)} & \mathbf{0}_{(N\times N)} \\ \mathbf{1}_{(1\times N)} & \mathbf{1}_{(1\times 1)} & \mathbf{0}_{(1\times N)} \\ \mathbf{0}_{(N\times N)} & \mathbf{1}_{(N\times 1)} & \mathbf{A}^{t+1} \end{bmatrix}, \end{aligned} \tag{5}$$

where $\cup$ denotes the operation of merging two adjacency matrixes that have temporal relations. Analogous to the associative property of multiplication, the final merged adjacency matrix remains the same when merging the adjacency matrixes of three or more temporally adjacent matrices. The merging of the adjacency matrixes for three temporally adjacent matrices is denoted by Eq. (6),

$$\mathbf{A}^t_{x\to g} \cup \mathbf{A}^t_{g\to x} \cup \mathbf{A}^{t+1}_{x\to g} = \begin{bmatrix} \mathbf{A}^t & \mathbf{0}_{(N\times 1)} & \mathbf{0}_{(N\times N)} & \mathbf{0}_{(N\times N)} \\ \mathbf{1}_{(1\times N)} & \mathbf{1}_{(1\times 1)} & \mathbf{0}_{(1\times N)} & \mathbf{0}_{(1\times N)} \\ \mathbf{0}_{(N\times N)} & \mathbf{1}_{(N\times 1)} & \mathbf{A}^{t+1} & \mathbf{0}_{(N\times 1)} \\ \mathbf{0}_{(1\times N)} & \mathbf{0}_{(1\times 1)} & \mathbf{1}_{(1\times N)} & \mathbf{1}_{(1\times 1)} \end{bmatrix}. \tag{6}$$

Figure 3 illustrates the merging process for the three temporally adjacent matrixes, where the merged spatial-temporal adjacency matrix in Fig. 3 corresponds to the description in Eq. (6). For a group video that comprises T frame images, the spatial-temporal emotional relationship can be represented by Eq. (7),

$$\mathbf{A}_f = \mathbf{A}^1_{x\to g} \cup \mathbf{A}^1_{g\to x} \cup \mathbf{A}^2_{x\to g} \cup \cdots \cup \mathbf{A}^T_{x\to g}, \tag{7}$$

where $\mathbf{A}_f$ denotes the spatial-temporal adjacency matrix, which characterizes the emotional relationships within a group on the spatial and temporal dimensions. Since in the adjacency matrix $\mathbf{A}_f$ the temporal relation of emotions is described by serially connecting nodes at different moments in time, the graph data based on adjacency matrix $\mathbf{A}_f$ is named as the spatial-temporal emotional graph with serial pattern.

For the pseudo (group) nodes introduced in the spatial-temporal emotional graph, the emotional feature matrix of a group at time t can be denoted as $\mathbf{X}^t_f = [\mathbf{x}^t_1, \mathbf{x}^t_2, \ldots, \mathbf{x}^t_N, \mathbf{g}^t]$, where the subscript f indicates that it is based on facial information. The emotional feature matrixes of nodes across all moments are represented as $\mathbf{X}_f = \left[\mathbf{X}^1_f, \mathbf{X}^2_f, \ldots, \mathbf{X}^T_f\right]$. At the t-th moment, the initial feature of the group node can be calculated by averaging the emotional features of all individuals at that moment:

$$\mathbf{g}^t = \frac{1}{N}\sum_{i=1}^{N} \mathbf{x}^t_i. \tag{8}$$

The spatial-temporal graph with a serial pattern for characterizing the emotional relationship within group video can be represented by $\mathbf{G}_f = \{\mathbf{A}_f, \mathbf{X}_f\}$.

2) Spatial-temporal Graph in Parallel Pattern: In the serial pattern, the spatial-temporal graph indicates the emotional influence from the individual's nodes to a group node, and then the updated group node affects the individual's nodes in the subsequent moment. If the feature learning of a group node at a certain moment is biased, it would significantly impact the feature learning of nodes in subsequent moments. On the other hand, the emotional state of an individual or group at a certain moment can influence not only the emotional state of other individuals or groups in the next moment but also the emotional states across multiple subsequent moments. This suggests that the influence of emotions has multiple parallel pathways, rather than a sequential, serial pattern.

Figure 4 illustrates the emotional influences in a parallel pattern. We present a parallel pattern spatial-temporal graph to represent the complex emotional relation in the spatial and temporal, whose adjacency matrix can be defined by Eq. (9),

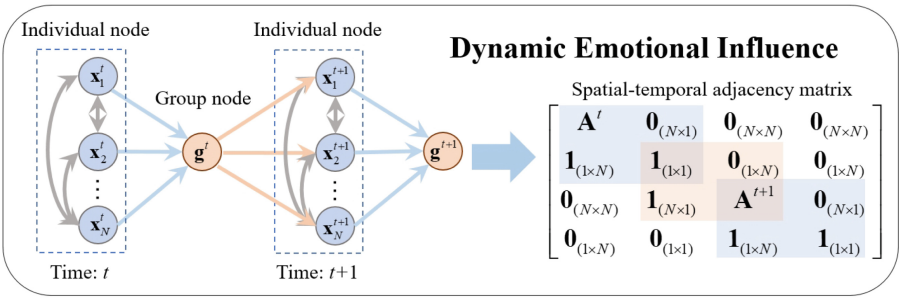


Fig. 3. The merged adjacency matrix for representing the spatial-temporal emotional relationship.

$$\widetilde{\mathbf{A}}_f = \varphi\left(\mathbf{A}_f\right), \tag{9}$$

where $\mathbf{A}_f$ is the spatial-temporal graph in a serial pattern defined, as defined in Eq. (7). $\widetilde{\mathbf{A}}_f$ denotes the adjacency matrix within the parallel pattern. $\varphi(\cdot)$ denotes the operation of replacing the element values of 0 in the lower triangular region with values of 1, and it ensures that the output matrix $\widetilde{\mathbf{A}}_f$ is a lower triangular matrix. The adjacency matrix $\widetilde{\mathbf{A}}_f$ indicates the emotional state of the nodes at the current time t is influenced by all nodes from time 1 to t. Furthermore, we employ a graph attention network to learn the attention weights of each edge, effectively focusing on the emotional information of significant nodes.

In summary, the emotional spatial-temporal graph with a parallel pattern can be represented as $\widetilde{\mathbf{G}}_f = \left\{\widetilde{\mathbf{A}}_f, \mathbf{X}_f\right\}$.

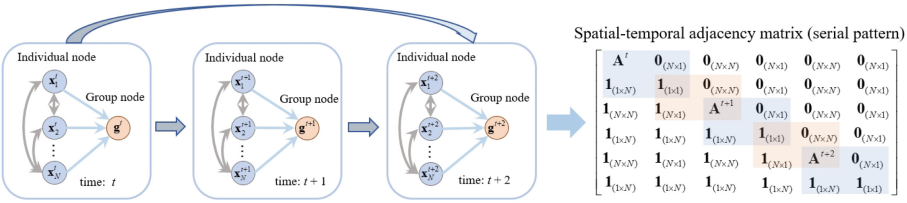


Fig. 4. The spatial-temporal emotional relationship in parallel mode.

3.4 Graph with Holistic Information

Holistic information from each frame image contains rich emotional cues and reflects the emotional state of a group at any moment. In this work, we also exploit the temporal correlations of holistic features with a graph-based model. We first construct a graph to represent the temporal correlation of the holistic information extracted from the group video. Specifically, we regard T consecutive

frames from a group video as T frame nodes and use the extracted holistic features (discussed in Sect. 3.2) as the nodal attributes. Similarly, we assume that the influence relationships among these T nodes follow a "parallel" mode in time. That is, the frame node at the t-th moment can affect the information of the frames at the moments of t+1, t+2, and so on, up to frame T. The temporal correlation of the T frame nodes can be represented by an adjacency matrix as shown in Eq. (10),

$$\mathbf{A}_h = \begin{bmatrix} a_{1,1} & \cdots & a_{1,T} \\ \vdots & \ddots & \vdots \\ a_{T,1} & \cdots & a_{T,T} \end{bmatrix}. \tag{10}$$

In the adjacency matrix $\mathbf{A}_h$, $a_{i,j}$ denotes the element of the i-th row and the j-th column; $a_{ij} = 1$, when index $i \leqslant j$; and $a_{ij} = 0$, when $i > j$. Equation (10) only simply describes there are edges (influence correlation) between T frame nodes. Furthermore, we employ the graph attention network [17] to learn the weights of edges to indicate the correlation degree from one node to another node temporally, thereby the topological relationship between the frame nodes can be dynamically characterized. In summary, the graph with holistic information can be represented as $\mathbf{G}_h = \{\mathbf{A}_h, \mathbf{X}_h\}$.

3.5 Graph Attention Network

We employ the graph attention network (GAT) [17] to separately exploit the constructed graph with facial and holistic information. By utilizing the GAT model, we can accurately learn the important relations between nodes, capture complex temporal and spatial dependencies with constructed graph data, and extract effective spatial-temporal features for GER. Using the graphs with facial and holistic information (introduced in Sects. 3.3 and 3.4) as inputs for two GAT models, the output results can be obtained by Eqs. (11) and (12),

$$\mathbf{X}'_f = \mathrm{GAT}_f\left(\mathbf{A}_f, \mathbf{X}_f\right), \tag{11}$$

$$\mathbf{X}'_h = \mathrm{GAT}_h\left(\mathbf{A}_h, \mathbf{X}_h\right), \tag{12}$$

where $\mathbf{X}'_f$ and $\mathbf{X}'_h$ denote the feature matrix of all nodes inferred through GAT models, respectively. In feature matrix $\mathbf{X}'_f$, the last row corresponds to the feature of the group node at time T (represented as $\mathbf{g}'^T$). Equation (11) is the process of exploiting the spatial-temporal graph constructed based on facial information, and the output feature $\mathbf{g}'^T$ is the spatial-temporal emotional representation which captures interactive information. We feed the feature $\mathbf{g}'^T$ into a fully connected layer to obtain GER results from the view of facial information (represented as $\hat{\mathbf{y}}_f$). Similarly, in the feature matrix $\mathbf{X}'_h$, the last row corresponds to the feature of the frame node at the moment T (represented as $\mathbf{h}'^T$) and it is the temporal representation of a group video. By feeding the feature $\mathbf{h}'^T$ into a fully connected layer, the GER results from the view of holistic information (represented as $\hat{\mathbf{y}}_h$) can be obtained.

3.6 Loss Function for Networks

By utilizing GAT models, we can exploit the constructed graph data and obtain temporal features of facial and holistic information, *i.e.*, $\mathbf{g}'^T$ and $\mathbf{h}'^T$. Furthermore, we use cross-entropy loss to supervise the parameter updating of both GAT models and the corresponding linear classifiers. To learn the spatial-temporal representation from facial information, we use the cross-entropy loss function with facial features as shown in Eq. (13),

$$L_f = -\log \frac{\exp\left(\mathbf{W}_{y_i}^{\mathrm{T}} \mathbf{g}'^T\right)}{\sum_{k=1}^{C} \exp\left(\mathbf{W}_k^{\mathrm{T}} \mathbf{g}'^T\right)}, \tag{13}$$

where $\mathbf{W}_k$ denotes the k-th classifier for the facial feature. To learn the temporal representation from holistic information of a group video, we use the cross-entropy loss function with holistic features expressed by Eq. (14),

$$L_h = -\log \frac{\exp\left(\mathbf{P}_{y_i}^{\mathrm{T}} \mathbf{h}'^T\right)}{\sum_{k=1}^{C} \exp\left(\mathbf{P}_k^{\mathrm{T}} \mathbf{h}'^T\right)}, \tag{14}$$

where $\mathbf{P}_k$ denotes the k-th classifier for the holistic feature. The proposed approach computes the loss L_f to supervise the parameter updating of the GAT_f model and the classifier $\mathbf{W}$, as well as computes the loss L_h to supervise the parameter updating of the GAT_h and the classifier $\mathbf{P}$.

3.7 Results Integration

We fuse the decisions based the facial and holistic information of a group video, *i.e.*, $\hat{\mathbf{y}}_f$ and $\hat{\mathbf{y}}_h$, to obtain a comprehensive recognition result for a group video. The final recognition result $\hat{\mathbf{y}}$ can be calculated as,

$$\begin{gathered} \hat{\mathbf{y}} = \alpha \hat{\mathbf{y}}_f + (1-\alpha)\hat{\mathbf{y}}_h, \\ \text{s.t. } 0 \leq \alpha \leq 1, \end{gathered} \tag{15}$$

where α is the fusion parameter for the two probability output vectors. We employ a grid search approach to obtain the suitable numerical value of α.

4 Experiments

To evaluate the performance of the proposed approach, extensive experiments were conducted on two widely used video-based GER datasets, including the Video Group Affect (VGAF) [15] and Group-level Emotion on the Crowded Videos (GECV) [14] datasets. VGAF is a large video dataset that contains 4183 collective videos. The videos in the VGAF dataset are divided into training, validation, and test sets, containing 2661, 766, and 756 videos respectively. Each video in the VGAF dataset has been manually annotated with one of three group emotion categories (*i.e.*, positive, neutral, and negative) by at least three

people. Since the test set of the VGAF dataset is unavailable to the public, the experiments only utilized the training and validation sets for performance evaluation. The GECV dataset contains 627 group videos. Due to legal and ethical reasons, some videos in the GECV dataset are not accessible, and only 408 videos are publicly available. Each video in the GECV dataset has also been manually annotated with one of the three group emotion categories: positive, neutral, or negative.

4.1 Implementation Details

Considering the computational efficiency of the proposed approach, for each frame image, we only use the top eight detected faces that have been detected by the MTCNN network [21]. The GAT models used in this work have two attention layers, and the dimension of the intermediate and output layers is 1024. A fully connected layer is set as a classifier for the output feature of the GAT model.

To train the GAT model and its classifier in the proposed method, we set the batch size of the constructed graphs with facial and holistic information to 64. The model's parameters are optimized using an Adam optimizer with a learning rate of 10^{-4}, and the maximum number of training epochs is set to 50. To fuse the results obtained by facial and holistic information, the fusion parameters α are set to 0.5 and 0.3 for the VGAF and GECV datasets, respectively. Empirically, we set the number of frames for each video sample to 20 and 15 in the VGAF dataset and GECV dataset, respectively.

All experiments were conducted on a Linux server with Intel Xeon CPU E5-2673 v4 2.30 GHz and GeForce GTX 2080Ti. We compared the proposed method with several state-of-the-art baseline methods that performed the same three-class classification [10,13,15,18] on the VGAF dataset. We compared the accuracy of our proposed method with the results reported in these research works. We also compared our method with several networks based on spatial-temporal convolutions, such as R(2+1), R3D, and MC3D, on the GECV dataset to evaluate the effectiveness of the proposed approach for video-based GER.

4.2 Comparison of Classification Performance

Tables 1 and 2 show the classification results of the proposed method and the baseline methods on the VGAF and GECV datasets, respectively. In this experiment, the proposed method employs the emotional spatial-temporal emotional graph with a parallel pattern as the input to the graph attention network. As there are different experimental setups present across the two datasets, the comparison results for each dataset are discussed separately.

Classification Results on VGAF Dataset: To ensure a fair comparison of the VGAF dataset, both the proposed method and the baseline methods were compared under the same emotional cues, and the overall accuracy results on the VGAF dataset were reported in Table 1. Specifically, Table 1 reports the

classification results obtained by using only the facial data source (titled "Facial level"), only the holistic data source (titled "Holistic level"), and both the facial and holistic data sources (titled "Facial + Holistic level"). The classification results in Table 1 show that the proposed method improves the best GER performance of using facial information by 64.36%. It demonstrates the effectiveness of the proposed method in exploiting the spatial-temporal emotional correlations. The comparison results of using facial and holistic data sources also indicate the superiority of the proposed method in capturing spatial-temporal information on constructed graphs and enhancing the performance of group emotion recognition. Furthermore, the comparison results show the performance discrepancy between the proposed method and the state-of-the-art methods in leveraging the holistic information of a group video. This is because the network employed for extracting holistic features in the proposed method is more efficient compared to these state-of-the-art methods. For instance, the compared methods [10,18] utilize the DenseNet [7] networks with 161 and 169 layers, respectively, while the proposed method employs a ResNet network [6] with 50 layers. These methods enhanced the effectiveness of feature representation, while their computational complexity is also much improved.

Table 1. Comparison of the overall classification accuracy on the VGAF dataset (in %). The symbol '-' indicates 'Not Reported'.

Method	Accuracy obtained by using emotional cues		
	Facial level	Holistic level	Facial + Holistic
Wang+ [18]	–	60.70	–
Pinto+ [13]	–	62.40	–
Sharma+ [15]	60.18	59.00	64.75
Liu+ [10]	63.71	**63.45**	68.02
Ours	**64.36**	57.44	**70.23**

Classification Results on GECV Datase: Considering there are only 408 video samples in the public version of the GECV dataset, we use the principle of ten-fold cross-validation to fully leverage the video samples for training and performance evaluation. We reproduced several spatial-temporal networks discussed in the literature [16] for comparison, including R(2+1)D, R3D, MC3D networks. We also compared with transformer architecture [11],*i.e.* Video Swin Transformer, for comparison. Additionally, the experiment reports the recognition performance of the proposed method when solely using facial or holistic information, *i.e.*, "Ours (Facial level)" and "Ours (Holistic level)". In Table 2, we show the average accuracies of the ten-fold cross-verification on single and overall categories in the columns marked by "Positive", "Neutral", "Negative" and "Overall". Comparison experiments show that the proposed method outperforms other methods in terms of recognition precision in individual categories as

well as overall accuracy. It indicates the proposed method is effective in exploiting the complex spatial and temporal relationships in constructed graphs. As the R(2+1)D, R3D, MC3, and Video Swin Transformer networks are the methods that exploit holistic information of a group video for GER, for a fair comparison, we compare the results of these methods with "Ours (Holistic level)". Comparison results show the proposed method is significantly more effective than the compared networks in leveraging holistic information. It also verifies the effectiveness of the GAT-based method proposed in this work in learning temporal features for GER.

Table 2. Comparison of average GER accuracy results obtained by the ten-fold cross-validation on the GECV dataset (%).

Method	Positive	Neutral	Negative	Overall
R(2+1)D [16]	93.68	89.15	81.25	89.01
R3D [16]	94.00	85.84	73.86	86.79
MC3 [16]	95.84	82.63	86.96	90.46
Video Swin Transformer [11]	88.89	76.92	70.00	80.49
Ours (Facial level)	91.17	87.58	84.59	88.73
Ours (Holistic level)	96.22	89.23	89.09	92.41
Ours	**96.65**	**90.56**	**89.75**	**92.90**

4.3 Ablation Study

In the proposed method, spatial-temporal emotional graphs in both the serial and parallel modes were constructed based on facial information. To investigate the effectiveness of the proposed method in representing emotional relationships within a group video and introducing pseudo (group) nodes into these graphs, we conduct two ablation studies: (1) the recognition accuracies obtained using spatial-temporal graphs in serial and parallel modes; (2) the impact of integrating pseudo (group) nodes in the parallel mode spatial-temporal graphs. Table 3 reports the GER results obtained by employing the GAT model to exploit three types of spatial-temporal emotional graphs. "Serial graph" and "Parallel graph" in Table 3 denotes the presented spatial-temporal emotional graphs with serial and parallel modes. "Parallel graph (w/o p)" in Table 3 involves constructing a parallel mode spatial-temporal graph without introducing pseudo nodes in the graph. Then the GAT model updates the attributes of the nodes in the graph, and the attributes of all nodes are averaged and fused for GER.

Table 3. Comparison of GER results obtained by exploiting several spatial-temporal emotional graphs. (in %)

Graphs	VGAF				GECV			
	Positive	Neutral	Negative	Overall	Positive	Neutral	Negative	Overall
Serial graph	60.60	57.50	45.65	55.88	88.30	81.85	73.82	82.85
Parallel graph (w/o p)	73.84	53.57	59.78	63.06	91.25	87.58	83.45	88.48
Parallel graph	71.19	62.50	55.98	64.36	91.17	87.58	84.59	88.73

The recognition results related to the "Serial graph" and "Parallel graph" in Table 3 demonstrate that using spatial-temporal graphs with a parallel pattern yields higher recognition accuracy. This suggests that the real emotional impact not only exists in serial paths where the group emotion influences the individual emotion of the next moment but also in paths with emotional impacts over multiple moments, implying the presence of multiple parallel paths. On the other hand, the comparison of with/without pseudo (group) nodes in the parallel mode spatial-temporal graphs reveals that the use of pseudo nodes enhances overall recognition accuracy. This indicates that introducing pseudo nodes into the emotional graph in this work effectively simulates the influence of individuals on the group and the feedback of the group on individual emotions. Consequently, it extracts useful interaction information and improves GER performance.

5 Conclusion

We propose a graph-based approach to leverage the facial and holistic information of a group video for GER. We first construct spatial-temporal graph data with facial information to effectively describe the emotional relationships within a group to the spatial and temporal dimensions. Then, we employ a graph attention network to exploit the constructed graph and dynamically model the emotional relationships and influences among individual and group nodes across the spatial-temporal dimension. By introducing the graph attention mechanism, the proposed method effectively focuses on the important nodes and captures interactive information to generate a more precise spatial-temporal representation for GER. Similarly, the proposed method also utilizes the GAT model to explore the temporal correlations of holistic features extracted from the frames of a group video and learns emotional temporal representations based on holistic information for accurate GER. Extensive experiments demonstrate that the proposed method learns effective spatial-temporal emotional features, and improves the accuracy of video-based GER from 68.02% to 70.23% on the VGAF dataset and from 90.46% to 92.90% on the GECV dataset.

Acknowledgments. This work was supported by National Natural Science Foundation of China (62173353), Science and Technology Program of Guangzhou, China (202007030011).

References

1. Barsade, S.G.: The ripple effect: emotional contagion and its influence on group behavior. Adm. Sci. Q. **47**(4), 644–675 (2002)
2. Dhall, A., Kaur, A., Goecke, R., Gedeon, T.: EmotiW 2018: audio-video, student engagement and group-level affect prediction. In: Proceedings of the 20th ACM International Conference on Multimodal Interaction, ICMI 2018, pp. 653–656, October 2018
3. Eyben, F., Wöllmer, M., Schuller, B.: Opensmile: the munich versatile and fast open-source audio feature extractor. In: Proceedings of the 18th ACM International Conference on Multimedia, MM 2010, pp. 1459–1462, October 2010
4. Fujii, K., Sugimura, D., Hamamoto, T.: Hierarchical group-level emotion recognition. IEEE Trans. Multimedia **23**, 3892–3906 (2021)
5. Guo, X., Polania, L., Zhu, B., Boncelet, C., Barner, K.: Graph neural networks for image understanding based on multiple cues: group emotion recognition and event recognition as use cases, pp. 2921–2930 (2020)
6. 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)
7. Huang, G., Liu, Z., Van Der Maaten, L., Weinberger, K.Q.: Densely connected convolutional networks. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2261–2269, July 2017
8. Kelly, J.R., Barsade, S.G.: Mood and emotions in small groups and work teams. Organ. Behav. Hum. Decis. Process. **86**(1), 99–130 (2001)
9. Lin, J., Gan, C., Han, S.: TSM: temporal shift module for efficient video understanding. In: 2019 IEEE/CVF International Conference on Computer Vision (ICCV), pp. 7082–7092, October 2019
10. Liu, C., Jiang, W., Wang, M., Tang, T.: Group level audio-video emotion recognition using hybrid networks. In: Proceedings of the 2020 International Conference on Multimodal Interaction, ICMI 2020, pp. 807–812, October 2020
11. Liu, Z., Ning, J., Cao, Y., Wei, Y., Zhang, Z., Lin, S., Hu, H.: Video swin transformer. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 3202–3211 (2022)
12. Parkhi, O.M., Vedaldi, A., Zisserman, A.: Deep Face Recognition. In: Proceedings of the British Machine Vision Conference 2015, pp. 41.1–41.12 (2015)
13. Pinto, J.R., et al.: Audiovisual classification of group emotion valence using activity recognition networks. In: 2020 IEEE 4th International Conference on Image Processing, Applications and Systems (IPAS), pp. 114–119, December 2020
14. Quach, K.G., Le, N., Duong, C.N., Jalata, I., Roy, K., Luu, K.: Non-volume preserving-based fusion to group-level emotion recognition on crowd videos. Pattern Recogn. **128**(C) (2022)
15. Sharma, G., Dhall, A., Cai, J.: Audio-visual automatic group affect analysis. IEEE Trans. Affect. Comput. **14**(2), 1056–1069 (2023)
16. Tran, D., Wang, H., Torresani, L., Ray, J., LeCun, Y., Paluri, M.: A closer look at spatiotemporal convolutions for action recognition. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 6450–6459, June 2018
17. Veličković, P., Cucurull, G., Casanova, A., Romero, A., Liò, P., Bengio, Y.: Graph attention networks, February 2018

18. Wang, Y., Wu, J., Heracleous, P., Wada, S., Kimura, R., Kurihara, S.: Implicit knowledge injectable cross attention audiovisual model for group emotion recognition. In: Proceedings of the 2020 International Conference on Multimodal Interaction, ICMI 2020, pp. 827–834, October 2020
19. Wang, Y., Zhou, S., Liu, Y., Wang, K., Fang, F., Qian, H.: ConGNN: context-consistent cross-graph neural network for group emotion recognition in the wild. Inf. Sci. **610**, 707–724 (2022)
20. Xie, H., et al.: Most important person-guided dual-branch cross-patch attention for group affect recognition, pp. 20598–20608 (2023)
21. Zhang, K., Zhang, Z., Li, Z., Qiao, Y.: Joint face detection and alignment using multitask cascaded convolutional networks. IEEE Sig. Process. Lett. **23**(10), 1499–1503 (2016)

Micro-expression Recognition Based on Dual-Stream Spatiotemporal Transformer

Yan Zhao[1], Xiaohua Huang[1,2,3(✉)], and Chuangao Tang[1]

[1] Oulu School, Nanjing Institute of Technology, Nanjing, China
{y00450220341,njit_tcg}@njit.edu.cn, xiaohuahwang@gmail.com
[2] Key Laboratory of Child Development and Learning Science (Southeast University), Ministry of Education, Southeast University, Nanjing, China
[3] Research Center for Learning Science, Southeast University, Nanjing, China

Abstract. Micro-expressions, imperceptible spontaneous facial movements reflecting underlying emotions, hold significant importance in emotion recognition. Due to their short duration and low intensity, micro-expression recognition (MER) remains challenging. The collection of micro-expressions poses difficulties due to their characteristics, leading to a scarcity of spontaneous micro-expression datasets. Furthermore, existing methods typically utilize only one type of input for MER, thus failing to fully exploit the limited micro-expression samples. To address these issues, we propose a new dual-stream spatiotemporal transformer network combining optical flow and magnified micro-expression, enabling to handle different types of information, thereby providing richer and more comprehensive representations. By simultaneously inputting both original micro-expression images and the corresponding optical flow change images into the dual-stream net-work, we obtain a diverse range of micro-expression information, consequently mitigating the impact of the scarcity of micro-expression datasets. Experimental evaluations conducted on three public datasets, namely SMIC, SAMM, and CASME II, demonstrate the superiority of our approach over other methods.

Keywords: Micro-expression recognition · Transformer encoder · Dual-stream spatiotemporal attention · Deep learning

1 Introduction

Facial expressions serve as outward reflections of individuals' inner worlds and play the most intuitive role in discerning others' emotions. However, in situations of danger or other contexts, most individuals adeptly employ false expressions to conceal genuine feelings, thus making it challenging to accurately comprehend their true thoughts. In such circumstances, determining individuals' actual emotional states through the observation of micro-expressions becomes crucial.

Micro-expressions are brief and imperceptible facial expressions that individuals attempt to suppress, disguise, or conceal their true inner emotions [1]. Being

A. Antonacopoulos et al. (Eds.): ICPR 2024, LNCS 15313, pp. 355–369, 2025.
https://doi.org/10.1007/978-3-031-78201-5_23

spontaneous [2], micro-expressions are uncontrollable and thus cannot be hidden by an individual, providing insights into one's genuine emotional state. Different from macro-expressions with longer duration, micro-expressions last for a very short period, ranging from 1/25 to 1/2 of a second [3], making them challenging to detect solely through visual observation, even for trained experts [4]. Therefore, it is necessary to develop a high-performance automatic micro-expression recognition algorithm in real-world.

According to the research of Ekman and Friesen [1], there are two types of micro-expressions that can reveal underlying emotions: "The time-reduced full affect micro displays (*i.e.*, micro-expressions) may well be those which the ego is not aware of, while the squelched micro displays may be those which the ego senses and interrupts in mid-performance." Therefore, current existing available spontaneous micro-expression databases [5–7] are collected according to two induction paradigms: requiring participants to completely suppress facial movements or to suppress facial movements upon becoming aware of facial actions. In contrast, posed micro-expression databases [8,9] collect data by instructing participants to mimic micro-expressions, which contradicts the spontaneity of micro-expressions. It can be observed that the collection and annotation of micro-expressions are extremely challenging, hence the scarcity of existing spontaneous micro-expression datasets. Therefore, the limited data volume makes it difficult to train models that can handle subtle changes in facial expressions and achieve outstanding performance.

Earlier micro-expression recognition methods primarily relied on handcrafted features, such as Local Binary Pattern (LBP) features. Recently, many works have developed deep learning based approaches for MER. Shao *et al.* [10] proposed I2Transformer architecture, which utilizes optical flow to extract global facial features and employs adversarial training strategies to remove identity interference. Subsequently, they utilize AU recognition as an auxiliary task to learn AU representations relevant to micro-expressions. Finally, they use transformers to process multiple AU representations and model the relationships between them to achieve MER. Takalkar *et al.* [11] utilized a dual attention network fed by upper face, lower face, and global face for MER. However, these methods only exploit one type, such as original images, motion magnified images, local face, or optical flow images, for MER, thus failing to fully exploit the limited micro-expression samples.

To address the aforementioned issues, we propose a dual-stream spatiotemporal transformer network, consisting of spatial attention and temporal attention networks capable of handling different types of information and temporal information, thereby providing richer and more comprehensive representations. Specifically, both magnified micro-expression frames and their corresponding optical flow change images are fed into the dual-stream spatial attention network, learning more discriminative information of micro-expressions. Subsequently, a simple but efficient temporal attention network is used to learn the temporal information from the obtained spatial features in micro-expression videos. Through these processes, we mitigate the impact of the scarcity of micro-expression datasets. The main contributions of this paper are as follows: (1) We

propose a dual-stream spatiotemporal attention module, where original micro-expression images and optical flow change images are fed into the dual-stream network. By concurrently learning features from both types of data, we obtain rich spatial information of micro-expressions. (2) By integrating CNN with the dual-stream spatial attention Transformer, our model is capable of processing subtle local information and global information, thereby obtaining information-rich representations of micro-expressions. (3) The superior performance of our proposed framework is validated on public micro-expression datasets, including SMIC, CASME II, and SAMM.

The rest of the paper is organized as follows. Section 2 will describe the related works about micro-expression recognition. We will describe our proposed method in Sect. 3. Section 4 will present the experiment results on three micro-expression databases. We will give a conclusion in Sect. 5.

2 Related Works

Micro-expression recognition involves classifying detected facial expression sequences based on their distinct features. Currently, there are three commonly used fundamental methods: Local Binary Pattern (LBP) feature-based [12–16], optical flow feature-based [10,17–19], and deep learning-based [20–25] micro-expression recognition methods. However, with the advancement of deep learning, both LBP feature-based and optical flow feature-based methods have also been combined with deep learning techniques [10,17,18].

LBP features have been widely used in the literature due to their simplicity in computation. By applying thresholding to the eight neighbors of each pixel and represent-ing the result with binary codes, LBP serves as a texture operator. However, LBP can only extract spatial information from single images and cannot directly capture temporal information from videos. To address this limitation, Zhao *et al.* [26] proposed Local Binary Pattern on Three Orthogonal Planes (LBP-TOP), which extracts LBP features from three orthogonal planes and cascades them to form new three-dimensional features. This approach enables the extraction of both temporal and spatial dimensions of information, thus achieving a transformation from two-dimensional to three-dimensional representations. Wang *et al.* [12] proposed LBP-SIP (Local Binary Pattern with Six Intersection Points) based on Zhao's work, which describes micro-expression features by taking the intersection points of LBP-TOP on three planes. The advantage of LBP-SIP over LBP-TOP is that it retains the ad-vantages of extracting spatiotemporal features while reducing feature redundancy and improving feature processing speed. Similarly, Huang *et al.* [13] also built upon Zhao's work and introduced Spatiotemporal Local Quantized Pattern (STCLQP), which incorporates sequence indicators as well as amplitude and direction information into spatiotemporal data. The improvement in STCLQP leads to richer feature information extraction and superior performance. Guo *et al.* [15] proposed a novel facial micro-expression recognition method called Extended Local Binary Pattern on Three Orthogonal Planes (ELBPTOP), which enhances recognition

accuracy and efficiency by analyzing local second-order information in video sequences. Additionally, the paper introduces the application of Whitened Principal Component Analysis for micro-expression recognition to obtain more compact and discriminative feature representations, significantly reducing computational costs.

Features based on optical flow: By calculating the relative motion information be-tween different frames, micro-muscle movements can be captured, which is useful for micro-expression recognition. Optical flow is used to describe the motion of brightness patterns in images, with the basic concept being to determine the distance traveled by the same object in different frames. As optical flow can capture temporal patterns between consecutive frames, it has been widely used in MER [19,27,28]. Verburg *et al.* [27] utilized Histogram of Oriented Optical Flow (HOOF) to en-code subtle changes in selected facial regions over time. Liong *et al.* [29] introduced another optical flow-based feature descriptor called Bi-Weighted Oriented Optical Flow (Bi-WOOF), which uses only two frames to represent a sequence of micro-expressions. Compared to HOOF, this method employs the magnitude and optical strain values of optical flow as a weighting scheme to highlight the importance of each optical flow, thereby reducing the influence of noise flows with smaller intensities. Recently, several researchers have combined optical flow features with or em-bedded into deep neural networks to further identify spatial patterns [18,28]. Li *et al.* [28] proposed an enhanced version of HOOF to reduce redundant dimensions. Zhang *et al.* [18] developed short and long range relation based spatio-temporal transformer based on long-term optical flow. However, there are certain drawbacks associated with using directional histogram of optical flow features: (1) If the input image quality is poor or contains noise, the directional histogram of optical flow features may be affected, leading to unstable recognition results. (2) Directional histogram of optical flow features mainly focuses on the motion of the selected facial region, but micro-expressions typically involve subtle changes across the entire face. Therefore, using only local information may not fully capture the features of micro-expressions. (3) The generalization ability of directional histogram of optical flow features may be limited. Its performance may be inconsistent across different faces, lighting conditions, and environments.

Recently, deep learning has become the most used method for micro-expression recognition [17,18,30–32]. Khor *et al.* [30] proposed a Rich Long-term Recurrent Convolutional Network (ELRCN), which first extracts features from each micro-expression frame using Convolutional Neural Networks (CNNs), and then processes the features using Long Short-Term Memory (LSTM) modules. Reddy *et al.* [31] utilized 3D CNNs to extract features from both spatial and temporal domains. However, these methods overlook the correlation between micro-expressions and Action Units (AUs). Additionally, facial identity information may interfere with the extraction of micro-expression features, thus limiting the accuracy of MER. To address these issues, Shao *et al.* [10] proposed a new method called I2Transformer, which enhances the accuracy of MER by learning invariant identity representations and modeling the relation-ship in transformer style.

This method utilizes optical flow to extract global facial features and removes identity interference through adversarial training strategies. Subsequently, AU recognition is utilized as an auxiliary task to learn AU representations relevant to micro-expressions. Finally, transformers are employed to process multiple AU representations and model the relationships between them to achieve MER. Fu *et al.* [17] introduced the Phase Driven Transformer (PDT), which utilizes optical flow features as input. PDT generates amplitude and phase information through two networks and combines them for network training. By incorporating image features into the frequency domain, PDT enhances the richness and diversity of features, enabling the model to extract more effective information and address the issue of unclear micro-expression features. Zhang *et al.* [18] proposed a novel spatio-temporal Transformer architecture, which is a purely transformer-based method for micro-expression recognition (*i.e.*, without using any convolutional networks). This architecture includes a spatial encoder for learning spatial patterns, a temporal aggregator for temporal dimension analysis, and a classification head. Liu *et al.* [19] pro-posed a new feature for spontaneous micro-expression recognition called Main Direction Mean Optical Flow (MDMO) feature. MDMO is a region-based, normalized statistical feature that considers local statistical motion information and its spatial locations, and micro-expression recognition is performed using a Support Vector Machine (SVM) classifier. Cen *et al.* [16] proposed a micro-expression recognition method based on a multitask facial action pattern learning framework and Joint Temporal Local Cubic Binary Pattern (Joint Temporal LCBP). This method explores the relationship between facial action units and emotional states by encoding temporal structures and subtle variations and utilizes regularization techniques to select meaningful feature subsets to improve recognition performance. As the scarcity of samples in micro-expression recognition poses a significant challenge, often leading to overfitting during the learning process and unsatisfactory recognition performance, some re-searchers have turned to transfer learning as a solution, which has shown promising results compared to previous approaches. Xia *et al.* [11] proposed a micro-expression recognition framework that utilizes macro-expression samples as guidance to train the micro-expression classifier. By extracting features from micro-expression and macro-expression samples, applying adversarial learning strategies, and using the triplet loss function, the micro-expression network can effectively capture shared features between them, thereby improving micro-expression recognition performance. Additionally, the incorporation of attention mechanisms has made excellent contributions to improving the accuracy of micro-expression recognition. Wang *et al.* [33] proposed a novel attention mechanism called Micro Attention, which is combined with residual networks to focus the network on facial regions with expression micro-movements. Moreover, to reduce the risk of overfitting when training deep networks on small datasets, the micro-attention unit is designed not to significantly increase parameters, and a simple yet effective transfer learning method is employed.

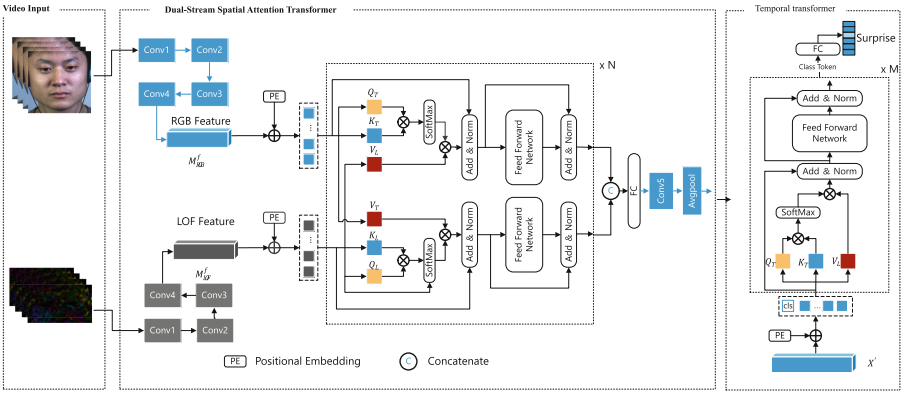


Fig. 1. Overview of Dual-stream Spatiotemporal Transformer.

3 Method

3.1 Overview

As shown in Fig. 1, the proposed method is based on a dual-stream spatiotemporal transformer network (DST) and primarily consists of a dual-stream spatial attention transformer (DSSAT) and a temporal transformer (TT). The model processes RGB images and optical flow images as input. The DSSAT extracts spatial facial features from each frame, while the TT generates discriminative feature representations by processing the spatial features from all frames. Finally, a fully connected (FC) network produces recognition results.

3.2 Dual-Stream Spatial Attention Transformer

Video Input: The DSSAT takes as input an RGB video of size $X_{RGB} \in \mathbb{R}^{T\times 3\times H\times W}$ and an optical flow video of the same size $X_{LOF} \in \mathbb{R}^{T\times 3\times H\times W}$, where $T = 17$. We extract a fixed-length sequence of 17 facial expression frames by cropping 8 preceding and 8 succeeding frames from the original video's vertex frame. The facial regions are cropped from these frames, and motion amplification techniques are applied to the video sequences, resulting in the input RGB video sequences. The optical flow video sequence computes long-term optical flow for the cropped 17 facial frames by calculating the optical flow between each frame and the initial frame among the 17 frames.

Convolutional Feature Extraction: We use four convolutional blocks to preliminarily extract feature maps $M_{RGB} \in \mathbb{R}^{C\times H^{'}\times W^{'}}$ and $M_{LOF} \in \mathbb{R}^{C\times H^{'}\times W^{'}}$ for each frame of RGB and optical flow images. These feature maps are then flatten into one-dimensional sequences $M^{f}_{RGB} \in \mathbb{R}^{D\times C}$ and $M^{f}_{LOF} \in \mathbb{R}^{D\times C}$, where $D = H^{'} \times W^{'}$. The input to the DSSAT is computed as follows:

$$z^{0}_{p,RGB} = m^{f}_{p,RGB} + e_{p,RGB}, \tag{1}$$

$$z_{p,LOF}^{0} = m_{p,LOF}^{f} + e_{p,LOF}, \tag{2}$$

where $e_{p,RGB} \in \mathbb{R}^{C}$ and $e_{p,LOF} \in \mathbb{R}^{C}$ are learnable position encodings that encode the spatial positions of the flattened images, and $p \in \{1, 2, \cdots, D\}$.

Calculation of Q, K, V: As shown in Fig. 1, the spatial encoder primarily consists of a cross-attention mechanism and fully connected layers. Three such spatial encoders form the DSSAT. In the l-th spatial encoder, $\mathbf{Q}$, $\mathbf{K}$, and $\mathbf{V}$ are obtained from the $(l-1)$-th spatial encoders $z_{p,RGB}^{l-1}$ and $z_{p,LOF}^{l-1}$.

In the cross-attention mechanism, each input feature has three weight matrices: query weight matrices $W_{Q,RGB}^{(l,k)}$,$W_{Q,LOF}^{(l,k)}$, key weight matrices $W_{K,RGB}^{(l,k)}$, $W_{K,LOF}^{(l,k)}$, and value weight matrices $W_{V,RGB}^{(l,k)}$,$W_{V,LOF}^{(l,k)}$. Therefore, $\mathbf{Q}$, $\mathbf{K}$, and $\mathbf{V}$ are defined as follows:

$$\begin{aligned} &\begin{bmatrix} Q_{p,RGB}^{(l,k)} & K_{p,RGB}^{(l,k)} & V_{p,RGB}^{(l,k)} \\ Q_{p,LOF}^{(l,k)} & K_{p,LOF}^{(l,k)} & V_{p,LOF}^{(l,k)} \end{bmatrix} = \\ &\begin{bmatrix} LN\left(z_{p,RGB}^{l-1}\right) & 0 \\ 0 & LN\left(z_{p,LOF}^{l-1}\right) \end{bmatrix} \begin{bmatrix} W_{Q,RGB}^{(l,k)} & W_{K,RGB}^{(l,k)} & W_{V,RGB}^{(l,k)} \\ W_{Q,LOF}^{(l,k)} & W_{K,LOF}^{(l,k)} & W_{V,LOF}^{(l,k)} \end{bmatrix}, \end{aligned} \tag{3}$$

where $\mathbf{Q}, \mathbf{K}, \mathbf{V} \in \mathbb{R}^{C'}$, $LN(\cdot)$ denotes the layer normalization, $k \in \{1, \cdots, S\}$ represents the index of the multi-head attention head, and S denotes the total number of multi-head attention heads. $C' = \frac{C}{S}$ represents the dimensionality of each attention head.

Dual-Stream Spatial Attention Transformer (DSSAT) Encoder: The DSSAT focuses on the fusion of features from different input images, namely RGB images and optical flow images. To more effectively learn the interaction features between these inputs, we simultaneously input $z_{p,RGB}^{0}$ and $z_{p,LOF}^{0}$ into the dual-stream fusion transformer. The output calculation formula is as follows:

$$\begin{cases} \hat{z}_{p,RGB}^{l} = DSSAT\left(LN\left(z_{p,RGB}^{l-1}, z_{p,LOF}^{l-1}\right)\right) + z_{p,LOF}^{l-1} \\ \hat{z}_{p,LOF}^{l} = DSSAT\left(LN\left(z_{p,LOF}^{l-1}, z_{p,RGB}^{l-1}\right)\right) + z_{p,RGB}^{l-1} \end{cases}, \tag{4}$$

$$\begin{cases} z_{p,RGB}^{l} = MLP\left(LN\left(\hat{z}_{p,RGB}^{l}\right)\right) + \hat{z}_{p,RGB}^{l} \\ z_{p,LOF}^{l} = MLP\left(LN\left(\hat{z}_{p,LOF}^{l}\right)\right) + \hat{z}_{p,LOF}^{l} \end{cases}, \tag{5}$$

where the $DSSAT(\cdot)$ function is formulated as follows:

$$DSSAT(LN(z_{p,RGB}^{l-1}, z_{p,LOF}^{l-1})) = softmax(\frac{Q_{p,RGB}^{(l,k)} K_{p,RGB}^{(l,k)^{T}}}{\sqrt{C'}}) V_{p,LOF}^{(l,k)}, \tag{6}$$

$$DSSAT(LN(z_{p,LOF}^{l-1}, z_{p,RGB}^{l-1})) = softmax(\frac{Q_{p,LOF}^{(l,k)} K_{p,LOF}^{(l,k)^{T}}}{\sqrt{C'}}) V_{p,RGB}^{(l,k)}, \tag{7}$$

and $\hat{z}_{p,LOF}^{l}$, and $\hat{z}_{p,RGB}^{l}$, are intermediate outputs of each dual-stream spatial encoder, $z_{p,LOF}^{l}$ and $z_{p,RGB}^{l}$ are the final outputs of each dual-stream spatial

encoder. And then, the two output derived from he last spatial encoders across RGB and LOF, *i.e.*, $z_{p,RGB}^{N}$ and $z_{p,LOF}^{N}$ are concatenated into one feature and fed into a fully connected layer, which is formulated as follows:

$$z_p^N = FC\left(concat\left(z_{p,RGB}^N, z_{p,LOF}^N\right)\right), \tag{8}$$

where $N = 3$, FC and *concat* represent the fully connected layer and concatenation operation, respectively, and z_p^N is the final output of the dual-stream spatial attention transformer. The D vectors z_p^N are concatenated to form the feature map $F \in \mathbb{R}^{C\times H'\times W'}$, where C, H', and W' represent the number of channels, height, and width, respectively. This process can be described as:

$$x_t' = GAP\left(g\left(F\right)\right), \tag{9}$$

where $g(\cdot)$ denotes a convolution operation and $GAP(\cdot)$ denoted global average pooling, respectively, and $t \in \{1, 2, \cdots, T\}$. The final output of the dual-stream spatial attention transformer is denoted as $X^{'} \in \mathbb{R}^{T\times P}$, where $P = C \times H' \times W'$.

3.3 Temporal Transformer (TT)

In this paper, we drew inspiration from the work of Zhao *et al.* [34] for the Time Transformer module. The Time Transformer consists of three time encoders, each encoder composed of multi-head self-attention mechanism and feed-forward networks. Here, we simply described TT architecture.

Given the spatial feature of one micro-expression video $X^{'} \in \mathbb{R}^{T\times D}$, position embeddings are added, leading an input embedding $z_{t'}^0$. Unlike the Dual-Stream Spatial Attention Transformer, a special learnable vector $x_0^{'} \in \mathbb{R}^P$ was added at the first position of the sequence. Subsequently, the calculation of $\mathbf{Q}$, $\mathbf{K}$, and $\mathbf{V}$ in the Time Transformer differs from the Dual-Stream Spatial Attention Transformer. While the Dual-Stream Spatial Attention Transformer needs to simultaneously process two different inputs, the Time Transformer only needs to handle one type of data, which is the output of the Dual-Stream Spatial Attention Transformer. $\mathbf{Q}$, $\mathbf{K}$, and $\mathbf{V}$ is computed as:

$$Q_{t'}^{(l,k)} = W_Q^{(l,k)} LN\left(z_{t'}^{l-1}\right) \in \mathbb{R}^P, \tag{10}$$

$$K_{t'}^{(l,k)} = W_K^{(l,k)} LN\left(z_{t'}^{l-1}\right) \in \mathbb{R}^P, \tag{11}$$

$$V_{t'}^{(l,k)} = W_V^{(l,k)} LN\left(z_{t'}^{l-1}\right) \in \mathbb{R}^P. \tag{12}$$

After obtaining the spatial relationship embedding for each frame from the Dual-Stream Spatial Attention Transformer, it is passed to the Time Transformer encoder to further learn long-range temporal relationships. We input $z_{t'}^0$ into the Time Transformer encoder, and the calculation formula is as follows:

$$z'^{l}_{t'} = TT\left(LN\left(z_{t'}^{l-1}\right)\right) + z_{t'}^{l-1}, \tag{13}$$

$$z_{t'}^{l} = MLP\left(LN\left(z'^{l}_{t'}\right)\right) + z'^{l}_{t'}, \tag{14}$$

where $TT\left(LN\left(z_{t'}^{l-1}\right)\right) = softmax(\frac{Q_{t'}^{(l,k)} K_{t'}^{(l,k)^T}}{\sqrt{F'}})V_{t'}^{(l,k)}$.

Lastly, the final classifier embedding is obtained from the first layer of the Time Transformer output, which is the special learnable vector mentioned earlier.

4 Experimental Analysis

In this section, we perform the experiments on SMIC [5], CASME II [7], and SAMM [6] databases for evaluating of the performance of our proposed method. Furthermore, we conduct the comparison with the state-of-the-art methods.

4.1 Database Description

The SMIC database comprises micro-expression sequences captured by high-speed cameras from 16 subjects, totaling 164 sequences. The filming was conducted at a frame rate of 100 fps. Micro-expression segments in SMIC were categorized into three types: positive (51), negative (70), and surprise (43).

CASME II consists of 247 micro-expression sequences from 26 subjects, including categories such as happiness (33), surprise (25), disgust (60), suppression (27), and others (102). These categories were labeled based on Action Units (AUs), self-reports from participants, and stimulus video content. All subjects are Chinese, and the participants are limited to young individuals from China.

SAMM dataset comprises 159 micro-expression sequences from 32 subjects. It includes eight categories: contempt (12), disgust (9), fear (8), anger (57), sadness (6), happiness (26), surprise (15), and others (26).

4.2 Experimental Setting

Our model was trained on a NVIDIA GeForce RTX 3090 GPU equipped with 24 GB of memory using the open-source PyTorch platform. The SGD optimizer is employed to optimize the parameters. Initially, the model was pretrained on the DFEW dataset with a batch size of 32, an initial learning rate of 0.01, and a division by 10 every 40 epochs, for a total of 100 epochs. The CASMEII, SAMM, and SMIC databases were then fine-tuned using the pretrained model from the DFEW dataset with a batch size of 16, a learning rate of 0.01, and a total of 100 epochs. The number of self-attention heads K was set to 8.

Experiment Protocol: For a fair comparison, all the experiments on each data-base are conducted with the leave-one-subject-out cross-validation, where samples from one subject are held out as the testing set while all remaining samples are used for training.

Performance Metric: In the experiments conducted on the CASME II, SAMM, and SMIC public datasets, the unweight F1-score (UF1) and accuracy (Acc) are used to measure the performance of various methods.

Table 1. Ablation Study on CASME II Dataset, where TT is temporal transformer, MI and OF represent magnified image and optical flow image, respectively.

	Input	UF1	Acc (%)
baseline	MI	0.7519	75.81
baseline	OF	0.7652	78.73
DSSAT (N=1)	MI+OF	0.8119	82.66
DSSAT (N=2)	MI+OF	0.8235	83.06
DSSAT (N=3)	MI+OF	0.8391	84.07
DSSAT (N=1) +TT	MI+OF	0.8386	84.58
DSSAT (N=2) +TT	MI+OF	0.8289	83.87
DSSAT (N=3) +TT	MI+OF	0.8561	86.69

4.3 Ablation Studies

An ablation study was conducted on the CASME II database, focusing on the number of spatial encoder and the impact of temporal transformer. The spatial encoder is shown in Fig. 2. The baseline method employed an architecture based on ResNet18 and Transformer, with magnified images as input. Additionally, for DSSAT, global averaging pool and a fully connected layer were used to classify the micro-expression sample. The results of ablation study are reported in Table 1.

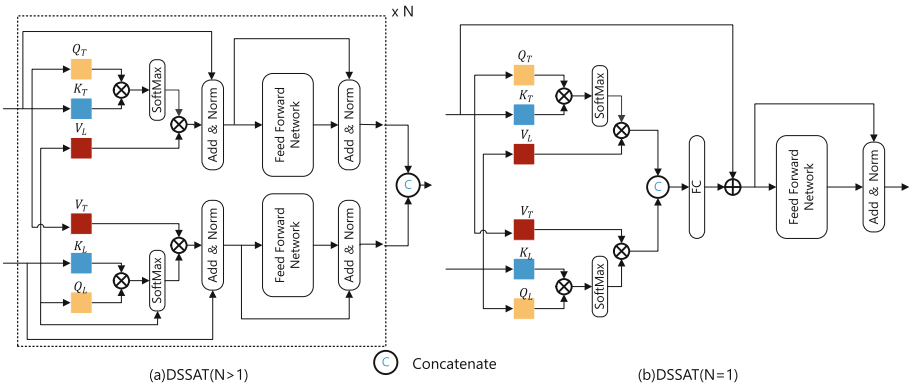


Fig. 2. The basic structure of DSSAT with N spatial encoders.

As shown in Table 1, the baseline architecture achieved a UF1 score of 0.7519 and an accuracy of 75.81%, when magnified image was used as input. This serves as a reference point for ablation studying our proposed model. When optical flow was incorporated into the baseline architecture, DSSAT (N = 1) improved upon the base-line, achieving a higher UF1 score of 0.8119 and accuracy of 82.66%. The improvement is achieved by an increase of 0.06 in terms of UF1 and 5.69%

Table 2. Performance comparison in terms of UF1 and accuracy on the CASME II database. The bold font indicates best performance.

Methods	UF1	Acc (%)
LBP-TOP [13]	0.424	46.46
LBP-SIP [13]	0.448	46.56
DiSTLBP-RIP [14]	–	64.78
STCLQP [13]	0.584	58.39
AU-GCN [21]	0.7047	74.24
SLSTT [18]	0.753	75.81
MMFRN [24]	–	63.51
GCL [25]	0.766	77.3
FeatRef [35]	–	62.85
I2transformer [10]	–	74.26
Ours	**0.8561**	**86.69**

Table 3. Performance comparison in terms of UF1 and accuracy on the SAMM database. The bold font indicates best performance.

Methods	UF1	Acc (%)
AU-GCN [21]	0.7045	74.26
SLSTT [18]	0.64	72.39
GCL [25]	**0.765**	77.1
GEME [36]	0.5467	65.44
FeatRef [35]	–	60.13
I2transformer [10]	–	68.91
Ours	0.7203	**78.31**

in terms of recognition rate. Furthermore, when the DSSAT model enhanced by increasing the number of spatial encoders, there is a further improvement in UF1 score of 0.8391 and accuracy to 84.07%.

The combination of DSSAT with N=1 and TT yields an even higher UF1 score of 0.8386 and accuracy of 84.58%. Furthermore, when $N = 3$ and TT, the proposed model achieves the highest UF1 score of 0.8561, with a slight decrease in accuracy to 86.69%. The experimental results demonstrate that both the DSST and the addition of TT contribute to improved performance in terms of UF1 score and accuracy, compared to the baseline model. Additionally, increasing the spatial encoder in the DSSAT model also leads to performance enhancements.

4.4 Comparison with State-of-the-Art Techniques

Our model was compared with two categories of baseline methods. The first cate-gory includes handcrafted feature-based methods (LBP-TOP [26], LBP-SIP

Table 4. Performance comparison in terms of UF1 and accuracy on the SMIC-HS database. The bold font indicates best performance.

Methods	UF1	Acc (%)
LBP-TOP [13]	0.538	53.66
LBP-SIP [13]	0.449	44.51
DiSTLBP-RIP [14]	-	63.41
STCLQP [13]	0.638	64.02
SLSTT [18]	0.74	75
GCL [25]	0.756	**77.2**
FeatRef [35]	–	57.90
GEME [36]	0.768	74.4
Ours	**0.7768**	76.22

[12], DiS-TLBP-RIP [14], and STCLQP [13]), while the second category comprises recent state-of-the-art deep learning methods (AU-GCN [21], SLSTT [18], MMFRN [24], GCL [25], and I2transformer [10]). The performance comparison is reported in Tables 2, 3 and 4.

According to Tables 2, 3 and 4, it can be observed that our method performs the best overall. Specifically, our method outperforms early hand-crafted works, such as LBP-TOP [26], LBP-SIP [12], STLBP-IP [14], and STCLQP [13], by a significant margin, demonstrating that deep models have more advantages in extracting micro-expression features. Moreover, compared with most state-of-the-art deep learning-based methods, such as, AU-GCN [21], SLSTT [18], MMFRN [24], GCL [25] and I2transformer [10], our method also achieves better results on the CASMEII dataset.

On the SAMM dataset, our method achieves the best accuracy result, surpassing the second-place method by 4.05%. While our method may not be the top performer in terms of UF1 score, it remains competitive. On the SMIC dataset, our method achieves the highest UF1 score, surpassing the second-place method by 0.0088. The accuracy is also only 0.98% lower than the top performer.

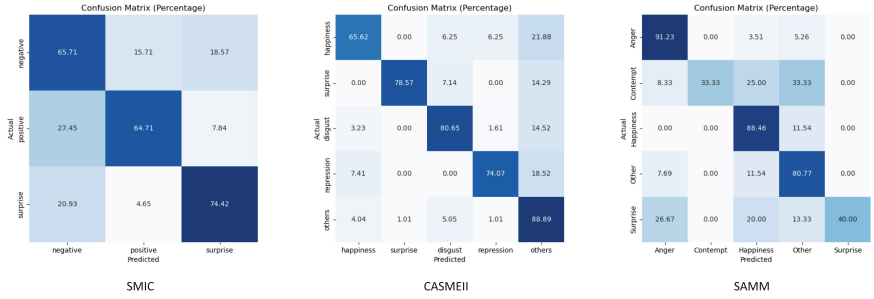


Fig. 3. The confusion matrices of our methods in the three databases.

Furthermore, Fig. 3 shows the confusion matrices of DST on three databases. As seen from Fig. 3, the confusion matrix in SMIC database indicates that the DST per-forms relatively well in recognizing surprise expressions, achieving an accuracy of 74.42%. For the CASME II database, the DST performs in others and disgust expression, while it struggles with recognizing happiness, achieving only a 65.62%. For the SAMM database, the DST performs reasonably well in recognizing anger expression and happiness expression. However, it struggles with recognizing contempt expressions, achieving only a 33.33% accuracy, and confuses them with other expressions. Moreover, there are notable challenges in accurately classifying surprise expression.

Overall, our proposed DST exhibits the promising performance on three databases comparing with several feature-engineering descriptors and deep learning architectures. Additionally, our proposed method performs well in recognizing certain expressions across the three databases.

5 Conclusion

In this paper, we propose a dual-stream spatiotemporal transformer network to process different types of information, thereby providing a richer and more comprehensive representation. We simultaneously input raw micro-expression images and optcal flow images into the dual-stream network to obtain rich micro-expression information, thus mitigating the impact of the limited number of micro-expression datasets. Specifically, we first use four convolutional blocks to preprocess RGB images and optical flow-transformed images to obtain feature maps. These feature maps are then simultaneously input into the dual-stream spatial attention transformer. We then utilize a temporal transformer to learn temporal features and finally output the classification results. This allows us to fully utilize micro-expression datasets. Extensive experimental results demonstrate that compared to existing methods, our approach achieves higher MER accuracy.

Acknowledgements. This research was supported by National Natural Science Foundation of China (Grant No. 62076122), the research funding of NJIT (No. YKJ201982), Basic Science (Natural Science) research project of higher education institutions in Jiangsu Province (24KJA520003), and the Fundamental Research Funds for the Central Universities (No. 2242024k30027).

References

1. Ekman, P., Friesen, W.V.: Nonverbal leakage and clues to deception. Psychiatry **32**(1), 88–106 (1969)
2. Ekman, P., Friesen, W.V.: Constants across cultures in the face and emotion. J. Personality Soc. Psychol. **17**(2), 124 (1971)
3. Ekman, P.: Lie catching and microexpressions. Philos. Decept. **1**(2), 5 (2009)

4. Frank, M., Herbasz, M., Sinuk, K., Keller, A., Nolan, C.: I see how you feel: Training laypeople and professionals to recognize fleeting emotions. In: The Annual Meeting of the International Communication Association, pp. 1–35 (2009)
5. Li, X., Pfister, T., Huang, X., Zhao, G., Pietikäinen, M.: A spontaneous micro-expression database: inducement, collection and baseline. In: International Conference on Automatic Face & Gesture Recognition, pp. 1–6. IEEE (2013)
6. Davison, A.K., Lansley, C., Costen, N., Tan, K., Yap, M.H.: Samm: a spontaneous micro-facial movement dataset. IEEE Trans.n Affect. Comput. **9**(1), 116–129 (2016)
7. Yan, W.-J., et al.: Casme ii: an improved spontaneous micro-expression database and the baseline evaluation. PLoS ONE **9**(1), e86041 (2014)
8. Polikovsky, S., Kameda, Y., Ohta, Y.: Facial micro-expressions recognition using high speed camera and 3d-gradient descriptor
9. Shreve, M., Godavarthy, S., Goldgof, D., Sarkar, S.: Macro-and micro-expression spotting in long videos using spatio-temporal strain. In: International Conference on Automatic Face & Gesture Recognition, pp. 51–56. IEEE (2011)
10. Shao, Z., Li, F., Zhou, Y., Chen, H., Zhu, H., Yao, R.: Identity-invariant representation and transformer-style relation for micro-expression recognition. Appl. Intell. **53**(17), 19860–19871 (2023)
11. Xia, B., Wang, W., Wang, S., Chen, E.: Learning from macro-expression: a micro-expression recognition framework. In: ACM International Conference on Multimedia, pp. 2936–2944 (2020)
12. Wang, Y., See, J., Phan, R.C.W., Oh, Y.H.: Lbp with six intersection points: reducing redundant information in lbp-top for micro-expression recognition. In: Asian Conference on Computer Vision, pp. 525–537. Springer (2015)
13. Huang, X., Zhao, G., Hong, X., Zheng, W., Pietikäinen, M.: Spontaneous facial micro-expression analysis using spatiotemporal completed local quantized patterns. Neurocomputing **175**, 564–578 (2016)
14. Huang, X., Wang, S.-J., Liu, X., Zhao, G., Feng, X., Pietikäinen, M.: Discriminative spatiotemporal local binary pattern with revisited integral projection for spontaneous facial micro-expression recognition. IEEE Trans. Affect. Comput. **10**(1), 32–47 (2017)
15. Guo, C., Liang, J., Zhan, G., Liu, Z., Pietikäinen, M., Liu, L.: Extended local binary patterns for efficient and robust spontaneous facial micro-expression recognition. IEEE Access **7**, 174517–174530 (2019)
16. Cen, S., Yang, Yu., Yan, G., Ming, Yu., Guo, Y.: Multi-task facial activity patterns learning for micro-expression recognition using joint temporal local cube binary pattern. Sig. Process. Image Commun. **103**, 116616 (2022)
17. Xiaofeng, F., Wenbin, W., Omata, M.: Phase driven transformer for micro-expression recognition. Multimedia Tools Appl. **83**(9), 27527–27541 (2024)
18. Zhang, L., Hong, X., Arandjelović, O., Zhao, G.: Short and long range relation based spatio-temporal transformer for micro-expression recognition. IEEE Trans. Affect. Comput. **13**(4), 1973–1985 (2022)
19. Liu, Y.-J., Zhang, J.-K., Yan, W.-J., Wang, S.-J., Zhao, G., Xiaolan, F.: A main directional mean optical flow feature for spontaneous micro-expression recognition. IEEE Trans. Affect. Comput. **7**(4), 299–310 (2015)
20. Zheng, Y., Blasch, E.: Facial micro-expression recognition enhanced by score fusion and a hybrid model from convolutional lstm and vision transformer. Sensors **23**(12), 5650 (2023)

21. Lei, L., Chen, T., Li, S., Li, J.: Micro-expression recognition based on facial graph representation learning and facial action unit fusion. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 1571–1580 (2021)
22. Liu, Y., Li, Y., Yi, X., Zuojin, H., Zhang, H., Liu, Y.: Lightweight ViT model for micro-expression recognition enhanced by transfer learning. Front. Neurorobot. **16**, 922761 (2022)
23. Nguyen, X.B., Duong, C.N., Li, X., Gauch, S., Seo, H.S., Luu, K.: Micron-bert: bert-based facial micro-expression recognition. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 1482–1492, June 2023
24. Zhang, J., et al.: Motion magnification multi-feature relation network for facial microexpression recognition. Complex Intell. Syst. **8**(4), 3363–3376 (2022)
25. Lao, L., Li, Y., Liu, M.L., Xu, C., Cui, Z.: Temporal discriminative micro-expression recognition via graph contrastive learning. In: International Conference on Pattern Recognition, pp. 1033–1040 (2022)
26. Zhao, G., Pietikainen, M.: Dynamic texture recognition using local binary patterns with an application to facial expressions. IEEE Trans. Pattern Anal. Mach. Intell. **29**(6), 915–928 (2007)
27. Verburg, M., Menkovski, V.: Micro-expression detection in long videos using optical flow and recurrent neural networks. In: International Conference on Automatic Face & Gesture Recognition, pp. 1–6. IEEE (2019)
28. Li, Q., Zhan, S., Liangfeng, X., Congzhong, W.: Facial micro-expression recognition based on the fusion of deep learning and enhanced optical flow. Multimedia Tools Appl. **78**, 29307–29322 (2019)
29. Liong, S.T., See, J., Wong, K., Phan, R.C.W.: Less is more: micro-expression recognition from video using apex frame. Sig. Process. Image Commun. **62**, 82–92 (2018)
30. Khor, H.Q., See, J., Phan, R.C.W., Lin, W.: Enriched long-term recurrent convolutional network for facial micro-expression recognition. In: International Conference on Automatic Face & Gesture Recognition, pp. 667–674. IEEE (2018)
31. Reddy, S.P.T., Karri, S.T., Dubey, S.R., Mukherjee, S.: Spontaneous facial micro-expression recognition using 3d spatiotemporal convolutional neural networks. In: International Joint Conference on Neural Networks, pp. 1–8. IEEE (2019)
32. Mao, Q., Zhou, L., Zheng, W., Shao, X., Huang, X.: Objective class-based micro-expression recognition under partial occlusion via region-inspired relation reasoning network. IEEE Trans. Affect. Comput. **13**(4), 1998–2016 (2022)
33. Wang, C., Peng, M., Bi, T., Chen, T.: Micro-attention for micro-expression recognition. Neurocomputing **410**, 354–362 (2020)
34. Zhao, Z., Liu, Q.: Former-dfer: dynamic facial expression recognition transformer. In: ACM International Conference on Multimedia, pp. 1553–1561 (2021)
35. Zhou, L., Mao, Q., Huang, X., Zhang, F., Zhang, Z.: Feature refinement: an expression-specific feature learning and fusion method for micro-expression recognition. Pattern Recogn. **122**, 108275 (2022)
36. Nie, X., Takalkar, M.A., Duan, M., Zhang, H., Xu, M.: GEME: dual-stream multi-task GEnder-based micro-expression recognition. Neurocomputing **427**, 13–28 (2021)

HR-TRACK: An rPPG Method for Heartrate Monitoring Using Temporal Convolution Networks

Lokendra Birla, Sneha Shukla, Trishna Saikia(✉), and Puneet Gupta

Indian Institute of Technology Indore, Indore, India
{phd2101101006,phd2201101014,puneet}@iiti.ac.in

Abstract. The COVID-19 pandemic necessitates avoiding skin contact to minimize the spread of virus infection. It paves the way for an active surge in telehealthcare research. In this direction, Remote Photoplethysmography (rPPG) plays a crucial role in analyzing heart rate (HR) from non-contact face videos. Existing rPPG-based HR monitoring methods fail when face video duration is small and the video contains facial deformations. These issues are mitigated by our proposed method *HR-TRACK*, that is, rPPG method for *H*eart *R*ate moni*T*oring using tempo*RA*l *C*onvolution networ*K*. It improves HR monitoring by introducing a novel architecture formed by sequentially stacking two novel networks. The networks are inspired by the temporal convolution network (TCN) to model long temporal sequences effectively. Our first network automatically mitigates the noise induced by facial deformations and performs blind source separation to predict pulse signals. The instantaneous HR obtained from the pulse signal can be erroneous. Thus, our second network analyzes all the computed HR values and rectifies the erroneous HR, if any. The experimental results conducted on the publicly available datasets reveal that our proposed method outperforms the state-of-the-art methods. Furthermore, the results justify the utilization of both networks to improve HR monitoring.

Keywords: Heart Rate · Temporal Convolution Network · Remote Photoplethysmograph · COVID-19

1 Introduction

In light of the significant impact of COVID-19 on healthcare systems, there has been a growing demand for telehealthcare solutions that enable remote analysis of individuals using non-contact technologies. Traditional monitoring techniques reliant on sensor contact for extended durations need to be improved, especially in contexts where minimizing physical contact is crucial due to infection concerns. Moreover, such methods are impractical for monitoring individuals such as newborns, those with compromised skin integrity, or individuals engaged in activities like sleep or exercise [31]. As a response to these challenges, Remote

A. Antonacopoulos et al. (Eds.): ICPR 2024, LNCS 15313, pp. 370–385, 2025.
https://doi.org/10.1007/978-3-031-78201-5_24

photoplethysmography (rPPG), a contactless technique, has gained popularity [9,31]. This non-contact approach utilizes facial videos captured by camera sensors to facilitate heart rate (HR) monitoring, offering versatility and accessibility across various applications in telehealthcare [14,15]. Non-contact HR monitoring has found utility in diverse areas, including cardiac disease monitoring and chronic disease treatment therapies in telehealthcare [27], affective computing for micro-expression recognition [10], stress analysis [23], biometrics for Deepfake detection [6], and spoof detection [2].

Traditionally, the rPPG technique involves HR estimation by performing the following steps: (i) finding the region of interest (ROIs); (ii) estimating temporal signals from different ROIs; (iii) denoising and consolidating the temporal signals to get a pulse signal; and (iv) computing the HR from the pulse signal. These steps are performed in several ways. For instance, [35] shows first that rPPG can be measured remotely from an RGB video, and [34] used the green color variation to compute the temporal signals as the green color variation provides better vital signals than the red and blue color variations. In [28], HR is computed from an input video by consolidating the temporal signals using independent component (ICA) based blind source separation (BSS). In [1], motion variations are analyzed instead of color variations to extract the temporal signals, and rPPG is computed using the principal component analysis (PCA) based BSS. The non-rigid motion from different face regions employs the Robust Accurate Direct Independent Component Analysis (RADICAL) technique to compute the HR [24]. These methods effectively mitigate noise issues due to facial deformations, head movement, and lighting variations in long-duration clips. However, they encounter challenges when analyzing short-duration clips commonly used in HR monitoring. Short-duration clips necessitate continuous computation of instantaneous HR, making the process more prone to errors than HR estimation and limiting the availability of relevant healthcare information [30]. Despite extensive research in HR estimation, efforts in HR monitoring remain constrained due to the prevalent issue of short-duration clips, resulting in limited available literature on rPPG-based HR monitoring.

Recent endeavors have focused on adapting existing HR estimation methods designed for long-duration videos to facilitate HR monitoring [30]. However, these techniques often fail to provide instantaneous HR, which is critical for healthcare applications [21]. Furthermore, the effectiveness of HR monitoring is hindered by the necessity of analyzing short-duration videos. In these videos, small differences between successive instantaneous HRs indicate more accurate estimation due to the gradual nature of HR changes [19]. To this end, by leveraging this insight, a rule-based Bayesian tracking method has been proposed to improve HR estimation by considering previous neighboring instantaneous HRs [14]. It illustrates efforts to improve performance by integrating multiple neighboring HR measurements and developing automated rectification techniques. However, further advancements are necessary in this direction.

The recent evolution of Deep Learning (DL), known for its effectiveness across various domains, has also gained popularity in this research field [17,22,26].

Typically, sequential neural networks are employed for this purpose [25]. Nevertheless, such networks are incapable of modeling the long sequence of temporal features. Thus, the effectiveness of rPPG-based HR monitoring could be diminished by utilizing such DL-based networks [8]. Hence, a suitable DL network and a corresponding strategy are needed to address these limitations.

By considering all these challenges, this paper proposes a novel HR monitoring method, *HR-TRACK*, that is, rPPG method for *H*eart *R*ate moni*T*oring using tempo*RA*l *C*onvolution networ*K*. It improves HR monitoring by introducing a novel architecture formed by sequentially stacking two novel networks. Both networks are inspired by the temporal convolution network (TCN). Our first network automatically mitigates the noise from the temporal signals and consolidates them to compute the instantaneous HR corresponding to the short-duration video clips. Subsequently, the second network analyzes all the HR values computed from the short-duration clips and rectifies the erroneous HR, if any. Our primary research contributions are:

1. We proposed a novel HR monitoring network to automatically rectify the spurious instantaneous HR computed from a short-duration clip. To this end, our network utilizes the signal-to-noise ratio. Moreover, it is inspired by TCN for modeling the instantaneous HR sequence in a better way than that of a sequence network.
2. Our first novel network will automatically consolidate and denoise the temporal signals by utilizing the action units (AUs). Moreover, it performs the blind source separation to predict the clean pulse signal, unlike [20], wherein only temporal signal denoising is performed, and that too for a specific facial region.
3. Our experimental results on publicly available datasets reveal that the proposed method performs better than state-of-the-art rPPG-based methods.

This paper is organized in the following way. The proposed method *HR-TRACK* is discussed in the next section. The experimental results are analyzed in Sect. 3. Sections 4 and 5 cover the discussion and conclusions, respectively.

2 Proposed Method

The proposed remote HR monitoring method, *HR-TRACK*, is presented in this section. It first divides the input face video into several overlapping short-duration fixed-size video clips, and instantaneous HR is computed for each clip. To this end, *HR-TRACK* extracts several ROIs from the video clip and estimates the temporal signal corresponding to each ROI. Subsequently, we select those temporal signals that are least affected by facial deformations and provide them to our first network, rPPG-TCN, to estimate the pulse signal. The network is inspired by TCN architecture because it models the long temporal sequences in a better way than other sequence models [3]. It utilizes the action units (AUs) to denoise the selected temporal signals and performs blind source separation automatically to consolidate denoised temporal signals for computing

the pulse signal. After that, we compute the instantaneous HR of the extracted pulse signal. The instantaneous HR computed from each short-duration clip can be spurious when the clip contains facial deformation. Thus, we utilize the neighboring instantaneous HR to rectify the spurious instantaneous HR. To this end, we propose Monitoring Rectification TCN (MR-TCN) for rectification. The flow diagram of *HR-TRACK* is shown in Fig. 1, in which input videos are divided into several overlapping clips. For visual clarity, each clip is represented in different colors. These clips are then passed through the ROI detection process. Each ROI is utilized for temporal signal estimation. The top 20 temporal signals with the smallest standard deviation are applied to the rPPG-TCN network for pulse signal estimation. These pulse signals are then used to compute instantaneous HR, which is further applied to the MR-TCN network for rectification. In the end, we obtained the non-spurious instantaneous HR.

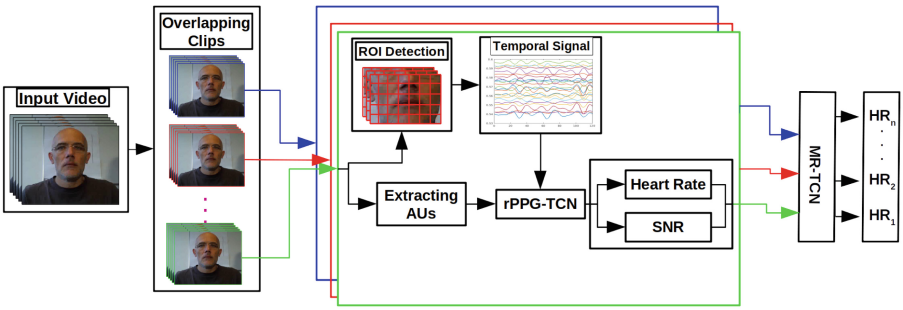


Fig. 1. Flow graph of the proposed method, *HR-TRACK*. We extract the multiple overlapping clips from the input video, and then ROIs are detected from each clip to compute the temporal signals. Subsequently, the top 20 temporal signals with the smallest standard deviation are passed to rPPG-TCN for pulse signal estimation and instantaneous HR estimated from the pulse signal. Finally, the instantaneous HR is passed through the MR-TCN to achieve rectified instantaneous HR.

2.1 Clips Extraction

The HR monitoring computes the instantaneous HR by analyzing several contiguous short-duration video clips instead of considering the full video, as in HR estimation. HR computation from long-duration video clips is avoided despite being resistant to some facial deformation because it does not lack the relevant information required for telehealthcare [29]. Unlike in [20], we have used overlapping clips for HR monitoring. Hence, we divide the input face video into overlapping short-duration video clips of equal duration. The HR monitoring will be performed in the subsequent stages by analyzing each clip. For dividing the input face video, we set the clip's duration to 4 s with a 2-second overlapping window (refer Sect. 3.3 for Parameter selection).

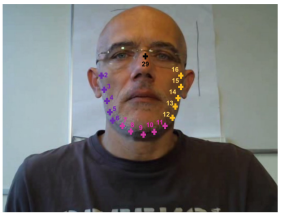


Fig. 2. Landmark points of input facial regions. (i) Points 7, 8, 9, 10, and 11 on the chin are highlighted in pink color. (ii) Points 2, 3, 4, 5, and 6 on the left cheek are shown in purple color. (iii) Points 12, 13, 14, 15, and 16 on the right cheek are indicated in yellow color. (iv) Point 29 on the nose is highlighted in black color. (Color figure online)

2.2 ROI Detection

The rPPG information is mainly present in the facial skin region. One of the ROI detection strategies employed in [21] is the Delaunay triangular method, in which the shape of the ROI is triangular. This method operates the landmark points as vertices to part the entire face region into multiple triangles. Each triangle is constructed by joining the three landmark points so that no other landmark point will be presented on their circumcircle. Different from the Delaunay triangular ROI detection method used in [20], we applied the square block ROI detection strategy in this paper. Hence, we first localize the face and extract several facial regions to compute the temporal signals. Our face localization requires estimating the facial landmark points that provide the facial boundary and outline the facial regions like eyes, nose, eyebrows, and mouth. This estimation is performed by applying CLNF Openface [38] to the first frame of the video clip. It provides 68 landmark points. Kindly note that we exclude the eyes and forehead regions because these regions provide an erroneous result due to eye-blinking and hairs, respectively. Thus, we utilize the remaining area below the eye's facial regions to define ROI. Specifically, this area is given by the convex hull of the lower face region's landmark points: i) 7, 8, 9, 10, and 11 of the chin; ii) 2, 3, 4, 5, and 6 of the left cheek; iii) 12, 13, 14, 15, and 16 of the right cheek; and, iv) 29 of the nose. Figure 2 depicts these landmark points on different facial regions. Furthermore, it is shown in [13] that facial deformation induces a large temporal variation at the facial boundary, which results in deteriorating the pulse signal. Hence, we remove the boundary pixels by utilizing the morphological erosion operations [13]. Moreover, it is shown in [14] that efficacy can be improved by considering several facial regions for temporal signal extraction. Thus, we divide the extracted region into multiple ROIs using the methodology proposed in [15]. For brevity, the complete region is divided into non-overlapping square blocks of size 10×10.

2.3 Temporal Signal Extraction

In this subsection, the temporal signal corresponding to each ROI is extracted. It is shown in [34] that the green color channel contains the strongest PPG signals because it is better absorbed by hemoglobin than the red color, and it penetrates human skin deeper than the blue color. Hence, we utilized the green color channel of an ROI to compute the temporal signals. That is, we compute the temporal signals by taking the average green color intensity of an ROI. Mathematically, temporal signal T_j corresponding to the ROI R_j, is given by:

$$T_j = \frac{\sum I_1^j(x,y)}{L(1)_j}, \frac{\sum I_2^j(x,y)}{L(2)_j}, ..., \frac{\sum I_k^j(x,y)}{L(k)_j} \tag{1}$$

where $(x, y) \in R^j$, the sum of intensities of the pixels in the green channel belonging to j^{th} ROI in the i^{th} frame is denoted by $\sum I_i^j(x, y)$. In an input clip, the total number of frames is k, and the total number of pixels in the j^{th} ROI of i^{th} frame is denoted by $L(i)_j$.

Along with the pulse signal, the extracted temporal signals contain respiratory signal information and the noises induced by facial deformation and illumination variations. The respiratory signal and other noises are mitigated from the temporal signals by applying a fourth-order Butterworth bandpass filter [36]. We set the range of the bandpass filter from 0.7 Hz (or 42/60) to 4 Hz (or 240/60) because the human heart beats at the rate of 42 to 240 beats per minute (BPM) [15]. Moreover, we apply the detrending filter to mitigate the noise generated by illumination variations [33]. Furthermore, it is shown in [13] that facial deformation induced by facial expressions deteriorates only some temporal signals. For instance, the temporal signals obtained from the mouth region are affected when the person is smiling. Hence, the standard deviation of such affected temporal signals is higher than the temporal signals not affected by the facial expressions [15]. Thus, we select the top 20 temporal signals having the smallest standard deviations and neglect the remaining signals for HR computation.

For better understanding, let us assume that $T_1, T_2, \cdots T_n$ are the temporal signals extracted from the face ROI region and $sd_1, sd_2, \cdots sd_n$ are their corresponding standard deviation values. The subscript n depicts the total number of temporal signals. If S_p is one of the selected temporal signals having the smallest standard deviation sd_p, then it can be represented as,

$$S_p = T_p \tag{2}$$

where, the index p is given by,

$$p = \underset{q \in (1,\cdots,n)}{argmin}\,(sd_q) \tag{3}$$

2.4 Extracting AUs

Facial expressions are usually studied in the literature in terms of action units (AUs), which analyze the face's different attributes [10]. Thus, we obtain the

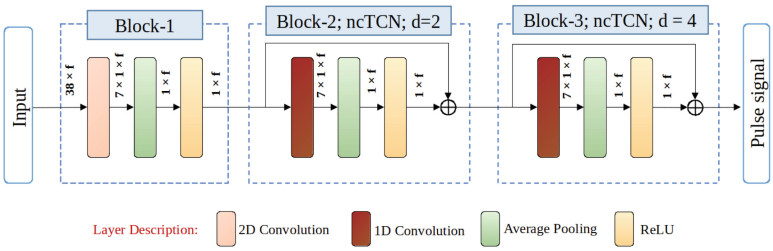


Fig. 3. The novel architecture of proposed rPPG-TCN. It requires the AUs and temporal signals computed from the input clip and provides the denoised pulse signal. d depicts the dilation rate.

AUs to mitigate the noise induced by facial expressions [20]. Following the AUs extraction method used in [20], we employ the CLNF Openface [38] to find the AUs from the face. The CLNF Openface takes an input clip and provides the 18 AUs for each frame of the clip. These AUs are provided in two ways: i) the presence of particular AU as binary numbers 0 and 1, and ii) the intensity of an AU ranges from 0 to 5. In this proposed method, we used the intensity of AUs for denoising pulse signals.

2.5 Pulse Extraction Using rPPG-TCN

The noise induced by facial deformations corrupts the temporal signals. In this subsection, we propose a novel network, rPPG-TCN, that denoises the temporal signals by mitigating such noise and automatically consolidates (or performs blind-source separation of) the denoised temporal signals for pulse extraction. It essentially requires the input clip's AUs and temporal signals to provide the denoised pulse signal. AUs are used to infer facial expression information. Moreover, our network is inspired by TCN architecture because it models the long sequence of information more effectively than sequential architectures [3]. The network architecture of rPPG-TCN is depicted in Fig. 3.

Our proposed rPPG-TCN network provides a denoised pulse signal using the temporal signals and AUs. Thus, the input size to rPPG-TCN is $38 \times f$, where f is the number of frames in the short-duration video clip, and 38 comes out by concatenating 18 AUs and 20 temporal signals. Our network is formed by sequentially stacking three blocks. Initially, the input is provided to the first block wherein 2D convolution is first performed using the seven filters, each of size 38×7. To perform the convolution operation, we used the same padding in the temporal direction, and in the other direction, we used no padding. The output size is $7 \times 1 \times f$, which is input to the average pooling layer to average the size in one direction and the Rectified Linear Unit (ReLU) activation layer to introduce non-linearity. After that, the resulting signal of size $1 \times f$ is passed to the second block, which is formed by taking inspiration from TCN architecture [3]. Specifically, we apply 1D convolutions on the resulting signal using the seven

filters, each having size 1×7 with the same padding. It results in a $7 \times 1 \times f$ signal, which is subsequently passed from the average pooling layer and ReLU activation layer to average the size in one direction and introduces non-linearity, respectively. We also add the skip connection between the input and output of the second block [3] and set its dilation rate to 2. After that, the resulting signal of size $1 \times f$ is passed to the third block. The third block is similar to the second block, with the difference that its dilation rate is set to 4. The output of our third block is the denoised pulse signal of size $1 \times f$. The details of the remaining parameters are provided in Sect. 3.2.

2.6 HR and SNR Estimation

This section computes the pulse signal's instantaneous HR and signal-to-noise ratio (SNR). The pulse signal mainly contains HR frequency and some noise. Thus, the frequency containing the maximum amplitude in the pulse spectrum corresponds to HR frequency [12]. Mathematically, the instantaneous HR, hr, is given by the below equation:

$$hr = f_{max} \times 60 \quad \text{where} \quad f_{max} = \underset{f}{\operatorname{argmax}}\, PS\,(f) \tag{4}$$

where f_{max} denotes the frequency of maximum amplitude in the pulse spectrum, and $PS\,[k]$ is the pulse spectrum's amplitude at frequency k^{th}. To compute the instantaneous HR, we multiply the frequency by 60 because the number of heartbeats in 60 s gives HR. Similarly, SNR [18] is given by:

$$SNR = \frac{\sum_{k=f_{max}-w_n}^{f_{max}+w_n} PS(k)}{\sum_{k=0.7}^{4} PS(k) - \sum_{k=f_{max}-w_n}^{f_{max}+w_n} PS(k)} \tag{5}$$

where w_n is the size of the neighboring frequency window chosen from [14]. The human heart beats 42–240 BPM, so we choose the frequency range of 0.7–4 Hz. Kindly note that the instantaneous HR computed from the short-duration video clip can be erroneous when the corresponding clip contains noise due to facial deformations. In such cases, the SNR will be high as opposed to the case when the pulse signal contains small noise.

2.7 HR Monitoring Using MR-TCN

This subsection proposes a novel network that analyzes the instantaneous HR and rectifies the erroneous instantaneous HR, if any. The network is referred to as MR-TCN, which is Monitoring Rectification TCN. It takes the instantaneous HR and SNR of short-duration clips. SNR is required to determine whether the instantaneous HR is erroneous or not. It is motivated by the observation that the SNR value is low when the clip contains facial deformations, resulting in erroneous instantaneous HR. Hence, our MR-TCN network employs the previous and future instantaneous HRs and SNR values to rectify the erroneous instantaneous HRs. Like our rPPG-TCN network, our MR-TCN network is inspired by

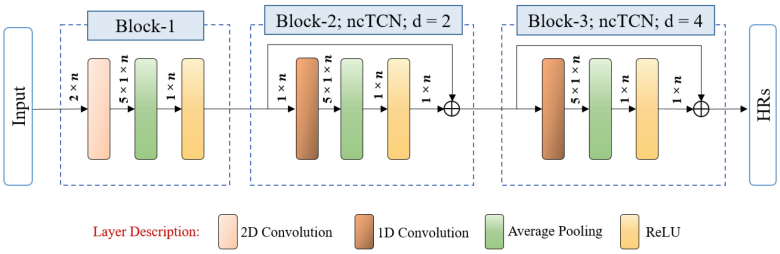


Fig. 4. The architecture of our proposed MR-TCN. It takes all the SNR values and instantaneous HRs to rectify the erroneous HRs, if any. d depicts the dilation rate.

TCN architecture to model the long temporal sequence effectively. The network architecture of MR-TCN is shown in Fig. 4.

The input size of our MR-TCN network is $2 \times n$, where n is the number of clips, and 2 comes up by concatenating instantaneous HR and SNR. The network is formed by sequentially stacking three blocks. The network's first block performs the 2D convolution using the five filters of size 2×5 with the same padding in the temporal direction and no padding in the other direction. The output has size $5 \times 1 \times n$, and it is passed to the average pooling layer applied in one direction to reduce the dimension. Then, the resulting signal is passed from a ReLU layer to introduce the non-linearity. This output is passed to the two sequential non-causal TCN (ncTCN) blocks. Both ncTCN block's input dimension is $1 \times n$, where 1D convolutions are applied using the five filters of size 1×5 with the same padding. The resulting output dimension is $5 \times 1 \times n$. Then, we apply average pooling in just one direction, and then we add ReLU activation. The final output has a size $1 \times n$. It corresponds to all the instantaneous HRs. We also add the skip connection between the input and output of each ncTCN block (inspired from [3]). The dilation rates are set to 2 and 4 for the second and third blocks of MR-TCN. The details of the remaining parameters are provided in Sect. 3.2.

3 Experimental Results

3.1 Dataset and Metrics

We performed the experimental evaluation of *HR-TRACK* on the publicly available datasets, UBFC-rPPG [4] and COHFACE [16] datasets. The COHFACE dataset contains 160 videos and the corresponding physiological signals recorded from 40 subjects. The duration and fps of these videos are 1 min and 20 fps, respectively. We used the 60% and 40% of training and testing ratios of subjects for comparative analysis using the COHFACE dataset, as suggested in [16]. Similarly, the UBFC-rPPG dataset contains 42 videos and the corresponding physiological signals recorded from 42 subjects. The duration and fps of these videos are 2 min and 30 fps, respectively. We used the 67% and 33% of training

Table 1. Comparative analysis of *HR-TRACK* with state-of-the-art methods

	UBFC-rPPG				COHFACE			
	SD*	MAE*	RMSE*	r	SD*	MAE*	RMSE*	r
ICA [28]	4.50	3.70	4.61	0.67	10.63	7.80	12.45	0.26
Chrominance-rPPG [7]	4.50	3.70	4.61	0.67	10.63	7.80	12.45	0.26
AHRE [11]	4.95	4.20	5.78	0.61	6.38	5.72	11.52	0.31
Fusion-EL [12]	4.20	3.71	4.52	0.73	8.09	7.14	9.43	0.57
RAHR [15]	4.50	3.70	4.61	0.67	10.63	7.80	12.45	0.26
MOMBAT [14]	3.38	3.50	4.01	0.85	6.14	5.89	7.92	0.62
Physnet [37]	3.85	3.63	5.29	0.94	7.90	8.59	11.60	0.36
META-rPPG [19]	4.50	3.70	4.61	0.67	10.63	7.80	12.45	0.26
Deepphys [5]	7.42	5.71	8.58	0.70	7.80	6.89	13.89	0.34
HR-CNN [32]	4.15	3.82	4.92	0.71	9.23	8.10	10.78	0.29
AND-rPPG [20]	3.21	2.67	4.07	0.96	4.53	3.82	5.10	0.79
HR-TRACK	**3.10**	**2.50**	**3.89**	**0.96**	**4.36**	**3.65**	**5.00**	**0.81**

and testing ratios of subjects for comparative analysis using the UBFC-rPPG dataset, as suggested in [4]. The dataset split is not mixed data from all participants. It is subject-independent data in which 60% and 40% of the subjects are taken as a training and testing set, respectively. These datasets provide the pulse signal as a ground truth acquired from the pulse oximeter during the video recording.

We compute the root mean square error (RMSE), the mean absolute error (MAE), the standard deviation (SD), and the Pearson's correlation coefficient (r) between the ground truth HR and estimated HR for performance evaluation [31]. We chose these metrics to provide a uniform comparison with the existing HR estimation methods, which utilized the same metrics for performance evaluation. Moreover, these metrics are mostly utilized for regression tasks. Since the HR estimation is one of the regression tasks in which the continuous output is estimated, thus, we employed these metrics in our proposed work. The lower value of these metrics indicates better network performance. Kindly note that all metrics are specified in bpm.

3.2 Implementation Details

Our proposed method *HR-TRACK* is implemented in Python using Pytorch. We perform experiments on the NVIDIA V100 GPU server and Intel Xeon Gold 6132 processor with 192 GB RAM. Initially, we train our rPPG-TCN network using negative Pearson correlation loss [29] with Adam optimizer. The corresponding learning rate is 0.001, the maximum number of epochs is 200, momentum is 0.9, dropout is 0.1, and weight decay is 0.0001. After training the rPPG-TCN network, we freeze its weight and train the MR-TCN network using cross-entropy

Table 2. Ablation study of our proposed method

	UBFC-rPPG				COHFACE			
	SD*	MAE*	RMSE*	r	SD*	MAE*	RMSE*	r
HR-TRACK	**3.10**	**2.50**	**3.89**	**0.96**	**4.36**	**3.65**	**5.00**	**0.81**
Without-SNR	3.16	2.57	3.94	0.94	4.40	3.72	5.03	0.80
Without-MRTCN	3.20	2.62	4.03	0.93	4.45	3.79	5.07	0.80
Denoised-TCN	3.50	2.81	4.31	0.86	4.90	4.01	5.46	0.72
Without-AUs	8.99	8.28	10.13	0.70	11.92	11.45	12.97	0.59
LSTM-exp	7.82	7.02	8.98	0.71	10.35	9.20	10.92	0.62
*WithoutTop*20	8.98	9.38	10.24	0.64	10.52	9.82	11.44	0.57

loss [39] with Adam optimizer. The corresponding learning rate, the maximum number of epochs, momentum, dropout, and weight decay parameters are set to 0.001, 100, 0.9, 0.1, and 0.0001, respectively. After training the MR-TCN network, we train both networks simultaneously for fine-tuning by unfreezing both networks. The fine-tuning is performed for 50 epochs. Since most of the samples in the existing datasets belong to the HR range of 70–90 BPM, but the normal HR range is 40–240 BPM. We mitigate this problem of unbalanced datasets by increasing the number of training samples. Thus, we utilize the data augmentation technique [20], employing linear interpolation to downsample and upsample the temporal signals. We downsample with a rate of 2 and 3 to generate higher HR samples while upsample the temporal signals to lower the HR.

3.3 Parameter Selection

Currently, it is a standard practice to use short-duration clips, typically 4-second clips, for HR monitoring [29]. If the clip size is less than 4 s, relevant HR information is discarded [29]. In contrast, if the clip size is larger than 4 s, then fewer clips are available for training, and the instantaneous HR information required for HR monitoring is lost. Therefore, we set the clip size to 4 s.

3.4 Comparative Evaluation

This subsection compares the proposed method with the existing state-of-the-art methods on the COHFACE and UBFC-rPPG datasets. The comparative performances are shown in Table 1. For fair comparative analysis, we rerun the publicly available codes[1] under the same experimental settings. That is, we follow the same testing protocol as used by [16,32] for fair comparative analysis. It can be analyzed from Table 1 that the proposed method, *HR-TRACK*, outperforms the state-of-the-art methods.

[1] https://github.com/lokendra7/rPPG-Publically.

The methods ICA [28], Chrominance-rPPG [7], and RAHR [15] perform poorly because these methods failed to mitigate the facial deformation-based noise from temporal signals. Also, the methods AHRE [11], Fusion-EL [12], and MOMBAT [14] demonstrate lower efficacy than the proposed method because our rPPG-TCN network automatically performs the denoising and blind source separation to mitigate several noises effectively. Similarly, our proposed method outperforms the Deepphys [5], PhysNet [37], and HR-CNN [32] because these methods failed to rectify the erroneous instantaneous HR and thereby provide inaccurate HR monitoring. We mitigate this issue by proposing an MR-TCN network. Likewise, META-rPPG [19] also exhibits lower performance than our method because it used the LSTM architecture, which is incompetent to model the long sequence of temporal information [3]. In contrast, we use TCN architecture to model long temporal sequences. Furthermore, our proposed method outperforms the AND-rPPG [20] because the proposed method effectively consolidates the different temporal signals automatically and also rectifies the erroneous instantaneous HR.

3.5 Ablation Study

We conduct several experiments to rigorously analyze the importance of AUs, rPPG-TCN, different ROIs, and MR-TCN. The corresponding results are shown in Table 2. These experiments are formed by changing or removing a subpart of the proposed method. The experiment details and the observations are as follows:

1. The experiment *Without-MRTCN* is performed by avoiding MR-TCN from our method. It can be observed from Table 2 that the experiment Without-MRTCN performs better than other state-of-the-art methods, shown in Table 1. However, its performance is lower than the proposed method, indicating that MR-TCN plays a crucial role in improving HR monitoring.
2. The experiment *Denoised-TCN* is formed by replacing the rPPG-TCN of our proposed with Denoised-TCN network and blind source separation employed in [20]. It can be observed from the table that our proposed method outperforms *Denoised-TCN*, indicating that our rPPG-TCN network has denoised the temporal signals and performed automatic blind-source separation more effectively than that proposed in [20].
3. In our proposed method, AUs provide relevant information about facial expressions. It is required to mitigate the noise due to facial expressions and thereby improve the HR computation. To experimentally justify the importance of AUs, we performed an experiment, *Without-AUs*, which is formed by avoiding the AUs in the rPPG-TCN of the proposed method. It can be analyzed from the table that our proposed method outperforms *Without-AUs*. It justifies the importance of AUs and advocates their utilization in HR monitoring.
4. Also, we replace the rPPG-TCN of our proposed method with LSTM architecture to form the experiment *LSTM-exp*. It is observed from the table that the proposed method's rPPG-TCN performed better than LSTM architecture

Table 3. Performance of *HR-TRACK* for different length clips

	UBFC-rPPG				COHFACE			
	SD*	MAE*	RMSE*	r	SD*	MAE*	RMSE*	r
2-Second	5.05	3.70	5.61	0.84	6.12	5.51	7.00	0.68
3-Second	4.10	3.01	4.92	0.90	5.09	4.47	6.20	0.72
4-Second	3.10	2.50	3.89	0.96	4.36	3.65	5.00	0.81
5-Second	3.05	2.37	3.84	0.97	4.25	3.59	4.99	0.82
6-Second	3.01	2.29	3.68	0.97	4.20	3.42	4.86	0.83
7-Second	3.11	2.28	3.56	0.96	4.28	3.51	4.82	0.82
8-Second	3.36	2.50	3.89	0.95	4.58	3.64	4.91	0.81
9-Second	3.42	2.74	4.01	0.94	4.81	3.84	5.05	0.79
10-Second	3.73	2.89	3.84	0.91	4.90	3.99	5.23	0.75

because sequential architecture failed to model the long sequence of temporal information.

5. In order to understand the importance of SNR, we design a method, *Without-SNR*, which is formed by avoiding the SNR in the MR-TCN network. It can be analyzed from the table that the performance decreases; it justifies the importance of SNR in MT-TCN.
6. To understand the selection of the top 20 temporal signals, we conducted a *WithoutTop*20 experiment in which we selected 20 random temporal signals to evaluate the proposed method, and it was found that performance decreased. The top 20 selection methods select the quality temporal signals and avoid noisy and spurious temporal signals.

4 Discussion

We have utilized the 4-second video clips to perform HR monitoring using the proposed method. We have also tested our proposed method for different time duration clips, and the corresponding results are shown in Table 3. It is analyzed from the table that if we increase the time duration of clips, then the performance of the proposed system increases for some time because long-duration clips can easily mitigate some noise. However, the performance decreases after some time because the number of clips required to train the model reduces. Also, it can be observed from the table that our best performance is achieved when 6-second video clips are used. Nevertheless, we set the clip size to 4 s because it is a widely accepted practice [29].

For rigorous analysis and a better understanding of the effectiveness of data augmentation in our proposed work, *HR-TRACK*, we conduct experiments without utilizing the data augmentation technique. Consequently, we obtained unsatisfactory results due to the imbalanced dataset used in our work because most

of the HR ranges lie between 70–90 bpm. In contrast, we achieve better results when data augmentation is applied in our work.

5 Conclusion

The rPPG technique plays a crucial role for doctors and patients in this COVID-19 pandemic to monitor the HR while avoiding the spread of the virus. Unfortunately, this technique is ineffective when the face contains deformations and short-duration face clips are utilized. Our proposed novel HR monitoring method, *HR-TRACK*, has effectively mitigated these limitations by sequentially stacking two novel networks inspired by the temporal convolution network (TCN). Our first novel network, rPPG-TCN, has automatically mitigated the facial expressions from the temporal signals and consolidated them to compute the instantaneous HR corresponding to the short-duration video clips. It has given some erroneous instantaneous HR when the corresponding clip contains facial deformation. Hence, all the computed instantaneous HR values have been provided to our second network, MR-TCN. The MR-TCN network has successfully rectified the erroneous instantaneous HR by analyzing all the instantaneous HR and SNR values. The experimental results were based on publicly available datasets, such as the UBFC-rPPG and COHFACE datasets. The experiments have revealed that the proposed method outperformed state-of-the-art methods. In the future, we intend to use the attention mechanism-based transformer architecture to improve HR monitoring. Additionally, we look forward to creating rPPG datasets for exceptional cases like dark skin tone.

Acknowledgment. The authors are thankful to all those researchers who have provided us the access to COHFACE and UBFC-rPPG datasets. This work of Trishna Saikia is partially supported by the Prime Minister's Research Fellowship (PMRF), the Ministry of Education, and the Government of India (2102743).

References

1. Balakrishnan, G., Durand, F., Guttag, J.: Detecting pulse from head motions in video. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 3430–3437 (2013)
2. Birla, L., Gupta, P.: PATRON: exploring respiratory signal derived from non-contact face videos for face anti-spoofing. Expert Syst. Appl. **187**, 115883 (2021)
3. Birla, L., Shukla, S., Gupta, A.K., Gupta, P.: ALPINE: improving remote heart rate estimation using contrastive learning. In: IEEE/CVF Winter Conference on Applications of Computer Vision, pp. 5029–5038 (2023)
4. Bobbia, S., Macwan, R., Benezeth, Y., Mansouri, A., Dubois, J.: Unsupervised skin tissue segmentation for remote photoplethysmography. Pattern Recogn. Lett. **124**, 82–90 (2019)
5. Chen, W., McDuff, D.: DeepPhys: video-based physiological measurement using convolutional attention networks. In: European Conference on Computer Vision, pp. 349–365 (2018)

6. Ciftci, U.A., Demir, I., Yin, L.: FakeCatcher: detection of synthetic portrait videos using biological signals. IEEE Trans. Pattern Anal. Mach. Intell. (2020). https://doi.org/10.1109/TPAMI.2020.3009287
7. De Haan, G., Jeanne, V.: Robust pulse rate from chrominance-based rPPG. IEEE Trans. Biomed. Eng. **60**(10), 2878–2886 (2013)
8. Gupta, A.K., Gupta, P., Rahtu, E.: FATALRead-fooling visual speech recognition models. Appl. Intell. **52**, 9001–9016 (2021)
9. Gupta, A.K., Kumar, R., Birla, L., Gupta, P.: RADIANT: better rPPG estimation using signal embeddings and transformer. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pp. 4976–4986 (2023)
10. Gupta, P.: MERASTC: micro-expression recognition using effective feature encodings and 2D convolutional neural network. IEEE Trans. Affect. Comput. **14**, 1431–1441 (2021)
11. Gupta, P., Bhowmick, B., Pal, A.: Accurate heart-rate estimation from face videos using quality-based fusion. In: IEEE International Conference on Image Processing, pp. 4132–4136 (2017)
12. Gupta, P., Bhowmick, B., Pal, A.: Serial fusion of eulerian and lagrangian approaches for accurate heart-rate estimation using face videos. In: International Conference of the IEEE Engineering in Medicine and Biology Society, pp. 2834–2837 (2017)
13. Gupta, P., Bhowmick, B., Pal, A.: Exploring the feasibility of face video based instantaneous heart-rate for micro-expression spotting. In: IEEE Conference on Computer Vision and Pattern Recognition Workshops, pp. 1316–1323 (2018)
14. Gupta, P., Bhowmick, B., Pal, A.: MOMBAT: heart rate monitoring from face video using pulse modeling and bayesian tracking. Comput. Biol. Med. **121**, 103813 (2020)
15. Gupta, P., Bhowmik, B., Pal, A.: Robust adaptive heart-rate monitoring using face videos. In: IEEE Winter Conference on Applications of Computer Vision, pp. 530–538 (2018)
16. Heusch, G., Anjos, A., Marcel, S.: A reproducible study on remote heart rate measurement. arXiv preprint arXiv:1709.00962 (2017)
17. Kuang, H., Ao, C., Ma, X., Liu, X.: Shuffle-rPPGNet: efficient network with global context for remote heart rate variability measurement. IEEE Sensors J. **23**, 15199–15209 (2023)
18. Lee, D., Kim, J., Kwon, S., Park, K.: Heart rate estimation from facial photoplethysmography during dynamic illuminance changes. In: International Conference of the IEEE Engineering in Medicine and Biology Society, pp. 2758–2761 (2015)
19. Lee, E., Chen, E., Lee, C.Y.: Meta-rPPG: remote heart rate estimation using a transductive meta-learner. In: European Conference on Computer Vision, pp. 392–409. Springer (2020)
20. Lokendra, B., Puneet, G.: AND-rPPG: a novel denoising-rPPG network for improving remote heart rate estimation. Comput. Biol. Med. **141**, 105146 (2021)
21. Macwan, R., Benezeth, Y., Mansouri, A.: Heart rate estimation using remote photoplethysmography with multi-objective optimization. Biomed. Signal Process. Control **49**, 24–33 (2019)
22. Mirabet-Herranz, N., Mallat, K., Dugelay, J.L.: Deep learning for remote heart rate estimation: a reproducible and optimal state-of-the-art framework. In: International Conference on Pattern Recognition, pp. 558–573 (2022)

23. Moghadam, M.C., Masoumi, E., Kendale, S., Bagherzadeh, N.: Predicting hypotension in the ICU using noninvasive physiological signals. Comput. Biol. Med. **129**, 104120 (2021)
24. Nooralishahi, P., Loo, C.K., Shiung, L.W.: Robust remote heart rate estimation from multiple asynchronous noisy channels using autoregressive model with kalman filter. Biomed. Signal Process. Control **47**, 366–379 (2019)
25. Nowara, E., McDuff, D., Veeraraghavan, A.: The benefit of distraction: denoising remote vitals measurements using inverse attention. arXiv preprint arXiv:2010.07770 (2020)
26. Odinaev, I., Wong, K.L., Chin, J.W., Goyal, R., Chan, T.T., So, R.H.: Robust heart rate variability measurement from facial videos. Bioengineering **10**(7), 851 (2023)
27. Parsi, A., Glavin, M., Jones, E., Byrne, D.: Prediction of paroxysmal atrial fibrillation using new heart rate variability features. Comput. Biol. Med. **133**, 104367 (2021)
28. Poh, M.Z., McDuff, D.J., Picard, R.W.: Non-contact, automated cardiac pulse measurements using video imaging and blind source separation. Opt. Express **18**, 10762–10774 (2010)
29. Qiu, Y., Liu, Y., Arteaga-Falconi, J., Dong, H., El Saddik, A.: EVM-CNN: real-time contactless heart rate estimation from facial video. IEEE Trans. Multimedia **21**, 1778–1787 (2018)
30. Rodriguez, A.M., Ramos-Castro, J.: Video pulse rate variability analysis in stationary and motion conditions. Biomed. Eng. Online **17**, 1–26 (2018)
31. Saikia, T., Birla, L., Gupta, A.K., Gupta, P.: HREADAI: heart rate estimation from face mask videos by consolidating eulerian and lagrangian approaches. IEEE Trans. Instrum. Meas. **73**, 1–11 (2023)
32. Spetlik, R., Franc, V., Matas, J.: Visual heart rate estimation with convolutional neural network. In: British Machine Vision Conference, pp. 3–6 (2018)
33. Tarvainen, M.P., Ranta-Aho, P.O., Karjalainen, P.A.: An advanced detrending method with application to HRV analysis. IEEE Trans. Biomed. Eng. **49**, 172–175 (2002)
34. Tasli, H.E., Gudi, A., den Uyl, M.: Remote PPG based vital sign measurement using adaptive facial regions. In: IEEE International Conference on Image Processing, pp. 1410–1414 (2014)
35. Verkruysse, W., Svaasand, L.O., Nelson, J.S.: Remote plethysmographic imaging using ambient light. Opt. Express **16**, 21434–21445 (2008)
36. Yang, M., Liu, J., Xiao, Y., Liao, H.: 14.4 nW fourth-order bandpass filter for biomedical applications. Electron. Lett. **46**, 973–974 (2010)
37. Yu, Z., Li, X., Zhao, G.: Remote photoplethysmograph signal measurement from facial videos using spatio-temporal networks. In: British Machine Vision Conference, p. 277 (2019)
38. Zadeh, A., Chong Lim, Y., Baltrusaitis, T., Morency, L.P.: Convolutional experts constrained local model for 3D facial landmark detection. In: IEEE International Conference on Computer Vision Workshops, pp. 2519–2528 (2017)
39. Zhang, Z., Sabuncu, M.R.: Generalized cross entropy loss for training deep neural networks with noisy labels. In: Neural Information Processing Systems (2018)

LLDif: Diffusion Models for Low-Light Facial Expression Recognition

Zhifeng Wang[1(✉)], Kaihao Zhang[2], and Ramesh Sankaranarayana[1]

[1] College of Engineering and Computer Science, Australian National University, Canberra, ACT, Australia
zhifengwang19@gmail.com

[2] School of Computer Science and Technology, Harbin Institute of Technology, Shenzhen, China

Abstract. This paper introduces LLDif, a novel diffusion-based facial expression recognition (FER) framework tailored for extremely low-light (LL) environments. Images captured under such conditions often suffer from low brightness and significantly reduced contrast, presenting challenges to conventional methods. These challenges include poor image quality that can significantly reduce the accuracy of emotion recognition. LLDif addresses these issues with a novel two-stage training process that combines a Label-aware CLIP (LA-CLIP), an embedding prior network (PNET), and a transformer-based network adept at handling the noise of low-light images. The first stage involves LA-CLIP generating a joint embedding prior distribution (EPD) to guide the LLformer in label recovery. In the second stage, the diffusion model (DM) refines the EPD inference, ultilising the compactness of EPD for precise predictions. Experimental evaluations on various LL-FER datasets have shown that LLDif achieves competitive performance, underscoring its potential to enhance FER applications in challenging lighting conditions.

Keywords: Low-Light · emotion recognition · diffusion model

1 Introduction

In the domain of computer vision, precisely identifying facial emotions presents a notable challenge, particularly in extremely low-light environments. Such environments can significantly impair the quality of captured images, leading to degraded visibility of facial features, which are crucial for precise emotion recognition. This degradation not only destroys the basic structure of the face but also introduces noise and distortion, further complicating the task for emotion recognition algorithms. In Fig. 1, the low-light image (LL) at the top shows a child's face that is shadowed and details are obscured, making it challenging to discern fine facial expressions. The histograms indicate that most pixel values are clustered toward the darker end of the spectrum, which suggests limited brightness and contrast in the image. In the normal-light image (CI) at the bottom, the child's face is clearly visible with good detail, essential for recognizing

A. Antonacopoulos et al. (Eds.): ICPR 2024, LNCS 15313, pp. 386–401, 2025.
https://doi.org/10.1007/978-3-031-78201-5_25

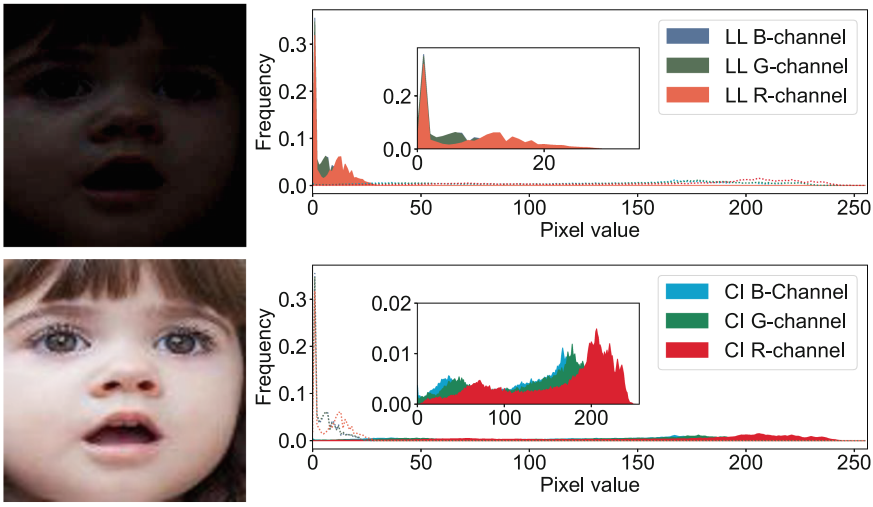


Fig. 1. Top: the low-light image (LL) shows a child's face that is shadowed and details are obscured, making it challenging to discern fine facial expressions. Bottom: In the normal-light image (CI) at the bottom, the child's face is clearly visible with good detail, essential for recognizing emotions.

emotions. The histograms show a more even distribution of pixel values across the spectrum, with higher frequencies in the mid to high ranges, indicating better brightness and contrast. Traditional FER methods [10,13,14,22,29] perform well under normal-light conditions; however, their effectiveness is considerably diminished in low-light scenarios due to the loss of subtle facial structures. There is a need for robust methodologies that can overcome the challenges posed by low brightness while maintaining high accuracy in emotion recognition. Currently, several approaches have been developed to tackle the challenge of learning from noisy data in the field of emotion recognition. RUL [25] proposes to improve facial expression recognition by weighing uncertainties based on the difficulty of samples to enhance performance in noisy environments. SCN [16] addresses uncertainties in facial expression recognition efforts by using a self-attention block to choose training samples and correcting uncertain labels by using a relabeling approach, thereby improving the learning process's dependability. However, both methods require relabeling the samples based on the samples' difficulties. EAC [26] addresses noisy labels by using flipped image consistency and selective features, preventing the model from relying on misleading features and thereby improving learning accuracy. However, when these techniques are used in low-light images, they encounter challenges. In particular, RUL [25] and EAC [26] are based on the assumption of minimal losses. In extremely low-light settings, where clear, fine facial details are lacking, these approaches might mistakenly equate challenging samples with noisy ones since both can display high loss values in the training of low-light images.

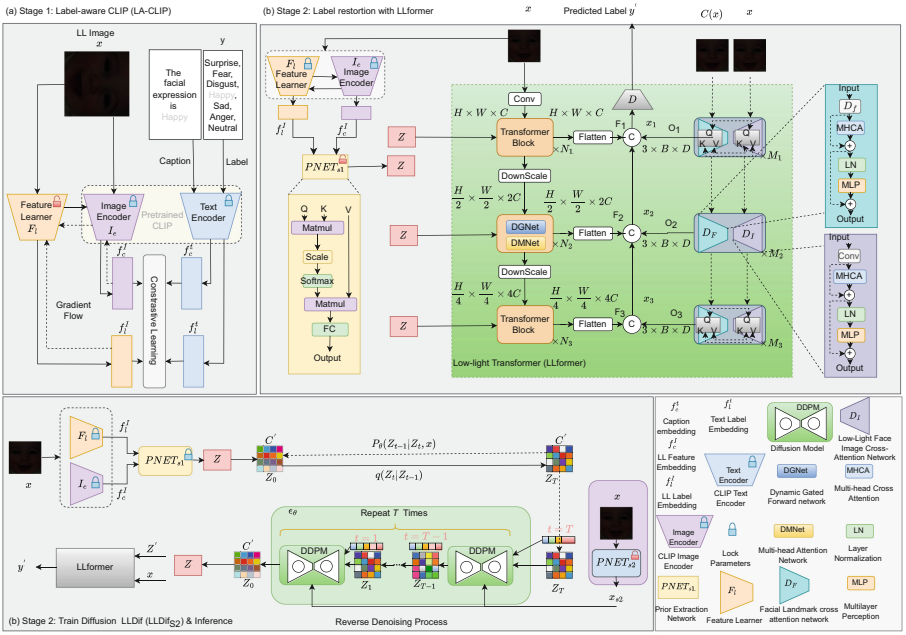


Fig. 2. The proposed LLDif framework, comprising Label-aware CLIP (LA-CLIP), LLformer, PNET, and a denoising network. LLDif employs a two-stage training method: (1) Initially, we apply LA-CLIP to process the low-light image alongside its image caption and label, producing a Joint Embedding Prior Distribution (EPD) Z. This EPD is then used to instruct the LLformer in label restoration. (2) During the second stage, the diffusion model's (DM) strong capabilities are employed to approximate the joint Embedding Prior Distribution (EPD) from PNET$_{s1}$. During the reverse process of the diffusion model, low-light images x are fed into PNET$_{s2}$ to derive a conditional vector x$_{s2}$.

To solve these issues, this paper proposes a novel method for handling noisy images in low-light conditions, departing from the conventional method of identifying noisy samples by their loss values. Instead, we introduce a distinctive approach centered on learning the joint distribution of noise labels and images via feature extraction and label restoration. We aim to create a diffusion-based network for FER that use the capabilities of diffusion models (DMs) for effective label restoration by aligning them with their related images. To achieve this, we present LLDif. Considering the transformer's capability to handle long-range pixel dependencies, we employ transformers as the foundational blocks of the LLDif architecture. We organize transformer blocks in a U-Net configuration to form the Low-Light Transformer (LLformer), which is aimed at extracting features at multiple levels. The LLformer comprises two parallel networks: the DTNet, tasked with extracting latent features from low-light images at various depths, and the DLNet, which focuses on identifying the similarities between

low-light images and facial landmarks. LLDif adopts a two-stage training approach: (1) In the first stage, as illustrated in Fig. 2 (a), we use LA-CLIP to process the low-light image along with its image caption and label, generating a Joint Embedding Prior Distribution (EPD) Z. This EPD is then utilized to guide the LLformer in label restoration. (2) In the second stage, shown in Fig. 2 (b), the diffusion model (DM) can be trained to deduce the accurate EPD directly from low-light images. Owing to the compactness of EPD Z, the DM can make highly accurate EPD predictions, achieving consistent high accuracy after only a few iterations.

This study offers several notable contributions, detailed as follows: 1) We introduce a innovative diffusion-based approach designed to address the challenges encountered in facial expression recognition, particularly those arising from diminished brightness and contrast in low-light conditions. 2) Our LLDif model harnesses the powerful distribution mapping capabilities of diffusion models (DMs) to generate an accurate embedding prior distribution (EPD), significantly enhancing the precision and reliability of FER results. This method stands out for its independence from the need to understand the dataset's uncertainty distribution, distinguishing it from prior approaches. 3) Extensive testing has demonstrated that LLDif achieves impressive performance in emotion recognition tasks across three low-light FER datasets, underscoring its effectiveness.

2 Related Work

Facial Expression Recognition. FER [24] focuses on enabling computers to interact with humans by identifying human facial expressions. In recent years, the accuracy of recognizing expressions under normal-light conditions has seen substantial improvements. Kollias *et al.* [4] introduces a CNN-RNN hybrid method that leverages multi-level visual features for dimensional emotion recognition. Zhao *et al.* [28] introduces Former-DFER, a dynamic transformer that combines spatial and temporal transformers to robustly capture facial features against occlusions and pose variations, achieving top performance on an emotion recognition dataset. The Expression Snippet Transformer (EST) [8] enhances video-based facial expression recognition by decomposing videos into expression snippets for detailed intra- and inter-snippet analysis, significantly outperforming conventional CNN-based approaches. Vazquez *et al.* [15] introduces a Transformer-based model, pre-trained on unlabeled ECG datasets and fine-tuned on the AMIGOS dataset, achieving top emotion recognition performance by leveraging attention mechanisms to emphasize relevant signal parts.

Diffusion Models. Diffusion models are now utilized across a wide range of tasks, including image enhancement for higher resolution, as mentioned by Shang *et al.* (2024) [12], and creative image modifications, as highlighted by Yang *et al.* (2023) [21]. Moreover, the latent features captured by diffusion models have proven beneficial for classification tasks such as image classification, as noted

by Han *et al.* (2022) [3], and for segmentation in medical imaging, as demonstrated by Wu *et al.* (2024) [20]. Zhang *et al.* [27] introduces a novel approach for editing single images using pre-trained diffusion models, combining model-based guidance with patch-based fine-tuning to prevent overfitting and enable high-resolution content creation and manipulation based on textual descriptions. Rahman *et al.* [11] presents a diffusion model-based approach for medical image segmentation that learns from collective expert insights to generate a variety of accurate segmentation masks, outperforming existing models in capturing natural variations and evaluated by a new metric aligned with clinical standards.

3 Methods

3.1 Label-Aware CLIP

The key idea of LA-CLIP is to train the feature learner F_l to output low-light features while simultaneously predicting the image's label. As summarized in Fig. 2 (a), the low-light feature embedding f_c^I is matched with the image's caption f_c^t. Moreover, the low-light label embedding f_l^I, predicted by the feature learner F_l, is aligned with the input label embedding f_l^t. This module helps to create embeddings that correlate visual features with textual annotations, which could be vital for low-light emotion recognition. It is designed to support the LLformer in label restoration, leveraging pre-trained models to guide the network in accurately predicting labels for low-light images.

As depicted in the yellow box of Fig. 2 during stage 2, $PNET_{s1}$ employs cross-attention layers to infer the Embedding Prior Distribution (EPD) Z. Following this extraction, DTNet leverages the EPD to aid in label recovery. Within DTNet, as shown in the same yellow box of Fig. 2, the architecture comprises DMNet and DGNet. We use the pre-trained LA-CLIP model to get the low-light feature embedding f_c^l and low-light label embedding f_l^I; these embeddings are then input into PNET$s1$. The output from PNET$s1$ is the EPD Z, denoted as $Z \in R^C$. This process is detailed in (Eq. 1):

$$Z = PNET_{S1}(F_l(x), I_e(x)). \tag{1}$$

Subsequently, Z is fed into the DTNet in Fig. 3, acting as adjustable parameters to support the process of label restoration, as detailed in Eq. (2).

$$F^{'} = W_1^l Z \circ LN(F) + W_2^l Z, \tag{2}$$

here, W represents the weights of a fully connected layer, LN denotes layer normalization and $\circ$ symbolizes element-wise multiplication. In DMNet Fig. 3 (b), we process the entire image to extract detailed information. The features F' are converted into three different vectors: key K, query Q, and value V, through a convolutional layer. These vectors are reshaped as Q to $R^{H''W'' \times C''}$, K to $R^{C'' \times H''W''}$, and V to $R^{H''W'' \times C''}$, making them compatible for subsequent operations. By multiplying Q and K, the model can identify which image regions

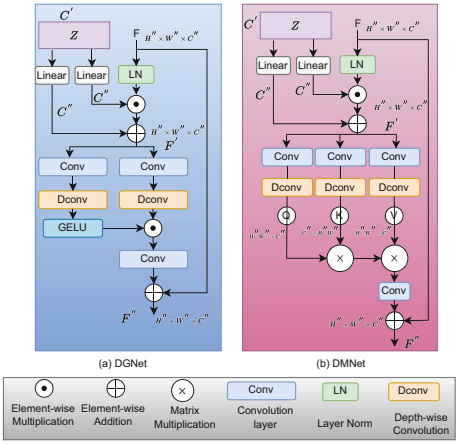


Fig. 3. The overview of DTNet, which consists of DGNet and DMNet.

to focus on, and generate an attention map $A \in R^{C'' \times C''}$. This operation in DMNet is depicted in the following Eq. (3):

$$F^{"} = W_c V \times softmax(K \times Q/\alpha) + F, \tag{3}$$

where α serves as a tunable parameter during the training phase. Following this, the DGNet focuses on extracting both local and neighboring features through aggregation. This is achieved by employing a small Convolution (1×1) to extract local features, and a larger Convolution (3×3) to collect information from adjacent pixels. Furthermore, a specialized gating mechanism is utilized to ensure only the most important information is captured. The entire process within DGNet is depicted in the following (Eq. (4)):

$$F^{"} = GELU(W_d^1 W_c^1 F^{'}) \circ W_d^2 W_c^2 F^{'} + F. \tag{4}$$

3.2 Dynamic Landmarks and Image Network (DLNet)

Within the DLNet, a cross window attention approach is utilized to process features from both 2D facial landmarks and related images taken in low-light conditions. We start by dividing the low-light image features, denoted as $X_{ll} \in R^{N \times D}$, into various distinct, non-overlapping windows $x_{ll} \in R^{M \times D}$. In parallel, features from facial landmarks, represented as $X_{fl} \in R^{C \times H \times W}$, are downscaled to align with the dimensions of these windows, yielding $x_{fl} \in R^{c \times h \times w}$, where the dimension c matches D and the production h and w equate to M. This setup enables the application of cross-attention between features of facial landmarks and low-light images, as depicted in (Eq. 7).

$$Q = x_{fl} w_Q, K = x_{ll} w_K, V = x_{ll} w_V, \tag{5}$$

$$O_i = Softmax(\frac{Q_i K_i^T}{\sqrt{d}} + b)V_i, i = 1, ..., N, \tag{6}$$

$$O = [O_1, O_2, ..., O_N]W_O, \tag{7}$$

where w_O, w_K, w_Q and w_V represent the weight matrices, and b denotes the corresponding positional bias.

This cross-attention mechanism is implemented on every window of the low-light image, termed as MHCA. The equations that describe the transformer encoder within LLDif are presented as follows (Eq. (9)):

$$X_{ll}^{'} = MHCA(X_{ll}) + X_{ll}, \tag{8}$$

$$X_{ll}^{"} = MLP(LN(X_{ll}^{'})) + X_{ll}^{'}, \tag{9}$$

the fusion of output features F from DTNet and O from DLNet is required to produce the combined multi-scale features x_1, x_2, and x_3. This involves concatenating the corresponding features: $x_1 = Concat(F_1, O_1)$, $x_2 = Concat(F_2, O_2)$, and $x_3 = Concat(F_3, O_3)$. Following this, the fused features X undergo additional processing through standard transformer blocks.

$$X = [x_1, x_2, x_3], \tag{10}$$

$$X^{'} = MSA(X) + X, \tag{11}$$

$$y^{'} = MLP(LN(X^{'})) + X^{'}, \tag{12}$$

where MSA denotes the self-attention blocks with multiple heads and LN refers to the layer normalization. The definition of the training loss is given as follows (Eq. (13)):

$$\mathcal{L}_{ce} = -\sum_{i=1}^{N}\sum_{c=1}^{M} y_{ic} \log(p_{ic}). \tag{13}$$

Our model is trained using the cross-entropy loss function, where M is the number of distinct classes, and N signifies the total count of samples. Here, y_{ic} indicates whether class c is the correct classification for observation i, and p_{ic} is the probability predicted by the model.

3.3 Diffusion Model for Label Restoration

In the second stage, as shown in Fig. 2 (b), the diffusion model's (DM) strong capabilities are employed to approximate the joint Embedding Prior Distribution (EPD). Initially, the pre-trained LA-CLIP and PNET_{S1} is used to acquire the EPD $Z \in R^C$. Following this, the diffusion technique is applied to Z, resulting in a generated sample $Z_T \in R^C$, as explained in (Eq. (14)):

$$q(Z_T|Z) = \mathcal{N}(Z_T; \sqrt{\bar{\alpha}_T}Z, (1 - \bar{\alpha}_T)I). \tag{14}$$

here, T represents the total count of diffusion steps. The variable α_t is defined as $1 - \beta_t$, and $\bar{\alpha}_T$ denotes the cumulative product of α_i for all steps from 0 to T. The term β_t is a predetermined hyper-parameter, while $\mathcal{N}(.)$ signifies the standard Gaussian distribution.

During the reverse process of the diffusion model, low-light images x are fed into PNET_{s2} to derive a conditional vector $x_{s2} \in R^C$ as outlined in Eq. (15).

$$x_{s2} = \text{PNET}_{s2}(x), \tag{15}$$

where PNET_{s2} includes convolutional layer, residual layer and linear layer, which will ensure the output's dimension of PNET_{s2} is same as PNET_{s1}.

The denoising network, represented as ϵ_θ, estimate the noise for each specific time step t. It processes the current noisy data $Z_t^{'}$, the time step t, and a conditional vector x_{s2}, which is obtained from the low-light image via the stage-two prior distribution network PNET_{s2}. The estimated noise, expressed as $\epsilon_\theta(\text{Concat}(Z_t^{'}, t, x_{s2}))$, is then utilized in the subsequent formula to determine the denoised data $Z_{t-1}^{'}$ for the upcoming step, as illustrated in Equation (16):

$$Z_{t-1}^{'} = \frac{1}{\sqrt{\alpha_t}}(Z_t^{'} - \epsilon_\theta(\text{Concat}(Z_t^{'}, t, x_{S2}))\frac{1-\alpha_t}{\sqrt{1-\alpha_t}}). \tag{16}$$

After T iterations, we get the final embedding prior distribution (EPD), symbolized as Z_0'. The stage-two prior distribution network (PNET_{s2}), together with the denoising network and the Low-Light Transformer (LLformer), are jointly optimized through the total loss function $\mathcal{L}_{total}$, as depicted in Equation (18).

$$\mathcal{L}_{kl} = \sum_{i=1}^{C} Z_{\text{norm}}(i) \log(\frac{Z_{\text{norm}}(i)}{\bar{Z}_{\text{norm}}(i)}), \tag{17}$$

$$\mathcal{L}_{total} = \mathcal{L}_{ce} + \mathcal{L}_{kl}. \tag{18}$$

In this formula, $Z_{\text{norm}}(i)$ and $\bar{Z}_{\text{norm}}(i)$ refer to the EPDs derived from LA-CLIP and LLDif_{S2}, respectively, both normalized through softmax. The term $\mathcal{L}_{kl}$ represents a form of the Kullback-Leibler divergence, computed over C dimensions. The total loss, $\mathcal{L}_{total}$, is formulated by adding the Kullback-Leibler divergence loss $\mathcal{L}_{kl}$ (Eq. 17) to the Cross-Entropy loss $\mathcal{L}_{ce}$ (Eq. 13). Since the EPD includes features from the low-light image and the corresponding emotion label encoded via a pretrained LA-CLIP model, LLDif's second stage (LLDif_{s2}) can provide accurate estimation for low-light image's label in a few steps. Notably, during the inference stage, LLDif doesn't need actual ground truth labels in the reverse diffusion process of DM.

4 Experiments

4.1 Datasets

LL-RAF-DB dataset includes 12,271 images in the training set and 3,068 images in the testing set, offering a robust basis for assessing FER algorithms

Table 1. Evaluation of accuracy (%) compared to SOTA FER methods on RAF-DB, KDEF and FERPlus.

RAF-DB		FERPlus		KDEF	
Methods	Acc. (%)	Methods	Acc. (%)	Methods	Acc. (%)
ARM [14]	90.42	DACL [2]	83.52	DACL [2]	88.61
POSTER++ [9]	92.21	POSTER++ [9]	86.46	POSTER++ [9]	94.44
RUL [25]	88.98	RUL [25]	85.00	RUL [25]	87.83
DAN [19]	89.70	DAN [19]	85.48	DAN [19]	88.77
SCN [16]	87.03	SCN [16]	83.11	SCN [16]	89.55
EAC [26]	90.35	EAC [26]	86.18	EAC [26]	72.32
MANet [29]	88.42	MANet [29]	85.49	MANet [29]	91.75
Ours	91.72	*Ours*	87.19	*Ours*	95.83

Table 2. Evaluation of accuracy (%) compared to SOTA FER methods on the LL-RAF-DB Dataset.

	DAN [19]	POSTER++ [9]	EAC [26]	MANet [29]	RUL [25]	SCN [16]	DACL [2]	*Ours*
Acc.(%)	79.27	80.76	78.72	78.45	77.57	75.20	75.68	82.26

under low-light conditions. Likewise, the RAF-DB dataset [7] includes 7 emotional categories and mirrors the testing and training configuration of LL-RAF-DB dataset. The expression distribution is consistent across both datasets.

LL-FERPlus dataset expands the scope to low-light conditions, presenting a comprehensive collection of 7,178 for testing and 28,709 images for training in low-light settings. The FERPlus dataset [1], an extension of the FER2013 dataset, is enriched with additional labels from ten different annotators and features the same quantity of training and testing images as the LL-FERPlus dataset.

LL-KDEF dataset contains 4,900 images captured under low-light conditions, taken from five unique angles. It comprises 3,920 images in the training set and 980 in the testing set. The KDEF dataset [6], with an identical total of 4,900 images, is a comprehensive collection in which each facial expression is photographed from five distinct viewpoints, ensuring a broad spectrum of clear visual information.

4.2 Implementation Details

We use Adobe Lightroom [5] to synthesize three benchmark low-light facial expression recognition (LL-FER) datasets following [?], simulating natural degraded image conditions by adjusting the exposure, white balance, highlights, and shadows and taking natural image statistics into consideration from normal-light FER images. Specifically, we start by generating three random variables, a, b, and c, each uniformly distributed between 0 and 1. These variables are

Fig. 4. Emotion distribution for samples in LL-RAF-DB dataset and RAF-DB.

Table 3. Evaluation of accuracy (%) compared to SOTA FER methods on the LL-FERPlus Dataset.

	DAN [19]	POSTER++ [9]	EAC [26]	MANet [29]	RUL [25]	SCN [16]	DACL [2]	*Ours*
Acc.(%)	80.97	81.44	80.46	80.34	79.35	74.95	77.05	82.25

then used to create parameters typical of those used in Adobe Lightroom software. The parameters include exposure $(-5 + a^2)$, highlights $(-20 \min\{b, 0.5\} + 5)$, shadows $(-20 \min\{c, 0.5\})$ and white balance $(-20(5 - 5a^2))$. The experimental setup utilized PyTorch for model training, which was carried out on a GTX-3090 GPU. For optimization, the Adam algorithm was chosen, with the training spanning 200 epochs. The adopted training settings specified an initial learning rate of 3.5×10^{-4}, a batch size of 64, and a weight decay parameter set to 1×10^{-4}.

4.3 Comparison with Other SOTA FER Methods

Comparison with Other Typical State-of-the-Art FER Methods. Table 1 offers a detailed evaluation of the accuracy of our proposed approach against the latest SOTA facial emotion recognition techniques [2,9,14,16,19,25,26,29] over three standard FER datasets: RAF-DB, KDEF and FERPlus. For RAF-DB, our method records a 91.72% accuracy, outperforming several well-established algorithms such as RUL [25], ARM [14], DAN [19], EAC [26], SCN [16] and MANet [29], and is closely matched with POSTER++ [9] which has a marginally higher accuracy of 92.21%. On the FERPlus dataset, the proposed method demonstrates an 87.19% accuracy, exceeding the accuracy of RUL [25] at 85.00%, POSTER++ [9] at 86.46%, EAC [26] at 86.18%, and MANet [29] at 85.49%, and the SCN [16] method has a lowest performance compared to the other methods. Within the KDEF dataset analysis, our proposed approach secures the top accuracy at 95.83%, showcasing a progress against other approaches, surpassing POSTER++ [9] at 94.44% and MANet [29] at 91.75%.

Table 4. Evaluation of accuracy (%) compared to SOTA FER methods on the LL-KDEF Dataset.

	DAN [19]	POSTER++ [9]	EAC [26]	MANet [29]	RUL [25]	SCN [16]	DACL [2]	*Ours*
Acc.(%)	82.03	88.93	43.53	83.13	83.99	77.50	86.69	92.97

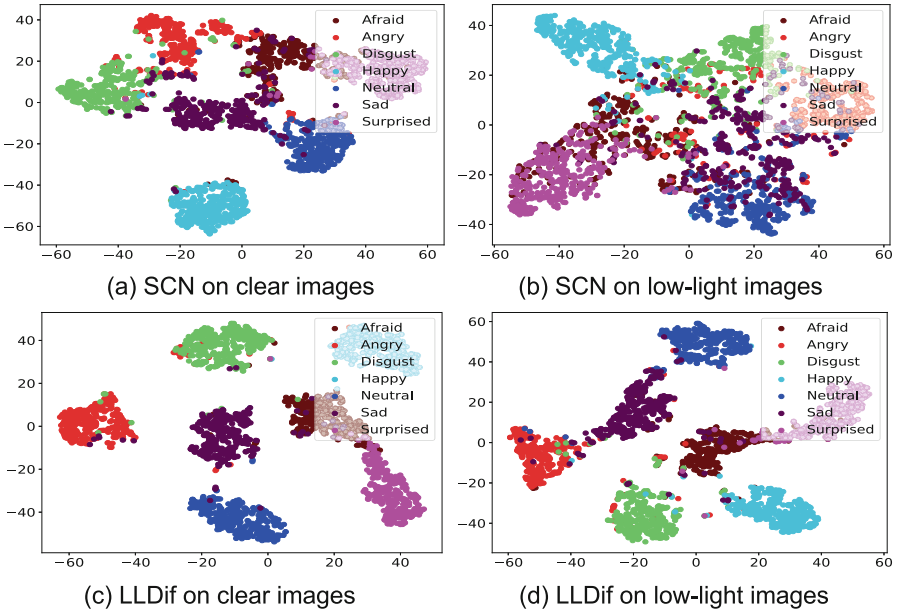


Fig. 5. The predicted feature visualised by t-SNE between our method and SCN.

Table 5. Key components in LLDif.

Method	Key components in LLDif				Acc.(%)
	Diffusion Model	$\mathcal{L}_{ce}$	$\mathcal{L}_{total}$	Insert Noise	
LLDif$_{S2}$-V1	✘	✔	✘	✘	89.46
LLDif$_{S2}$-V2	✔	✘	✔	✔	91.67
LLDif$_{S2}$-V3	✔	✔	✘	✘	92.16
LLDif$_{S2}$-V4 (Ours)	✔	✔	✘	✔	92.97

Overall, these results underscore the reliability of the proposed approach in handling facial expression recognition across various datasets.

Comparison with the Low-Light FER-Model. We compare our method with other SOTA methods on low-light images. Some samples are shown in Fig. 4. Accuracy comparisons between our model and other SOTA FER methods on the

LL-RAF-DB, LL-KDEF datasets, and LL-FERPlus are outlined in Tables 2, 3, and 4, respectively. The majority of the benchmarked models, including ARM [14], RUL [25], DAN [19], SCN [16], EAC [26], and MANet [29], are based on the ResNet-18 architecture. However, POSTER++ [9] stands out by adopting the Vision Transformer architecture. In contrast, our model introduces a novel approach by incorporating a 'Diffusion' backbone, moving away from the traditional ResNet-18 design. In the Table 2, the proposed method attains the highest accuracy of 82.26%, which is a notable enhancement over other methodologies. POSTER++ [9] registers the second highest accuracy with 80.76%, followed by DAN [19]at 79.27%. The EAC [26], MANet [29], RUL [25], SCN [16], and DACL [2] algorithms show a relative low accuracy from 75.20% to 78.72%. Table 3, which focuses on the LL-FERPlus Dataset, shows "Ours" with a leading accuracy of 82.25%, marginally surpassing POSTER++'s [9] 81.44%. In the Table 4, "Ours" shows the highest accuracy at 92.97%, which is significantly higher than the other methods listed. The second most accurate method is POSTER++ [9], with an accuracy of 88.93%. Other methods such as DAN [19], EAC [26], MANet [29], RUL [25], SCN [16], and DACL [2] present accuracies ranging from 43.53% to 86.69%. These results underscore the efficiency of the diffusion-based approach within the context of facial expression recognition systems under low-light conditions.

Feature Visualization. We used the t-SNE method to illustrate how models discern feature distributions. In contrast to Fig. 5 (a) and (b), where the SCN model has difficulty separating different emotion categories, especially in low-light conditions, our LLDif model exhibits effective expression recognition in both clear and degraded low-light images. This indicates that LLDif successfully captures key features crucial for distinguishing between various emotional expressions categories.

Visualization of Confidence Scores. We visualize the distribution of confidence scores for facial expression recognition methods on clear and low-light images in Fig. 6. For the baseline method [16], the mean confidence score for clear images is 0.41 and for low-light images is 0.34, with an overall accuracy of 0.435. The DAN method [19] shows a mean confidence score of 0.42 for clear images and 0.37 for low-light images, with an overall accuracy of 0.820. The POSTER++ method [9] has mean scores of 0.46 for clear images and 0.45 for low-light images, achieving an overall accuracy of 0.889. The proposed method exhibits a notably higher confidence level with mean scores of 0.57 for clear images and 0.53 for low-light images, corresponding to a high overall accuracy of 0.929. The proposed method not only shows the highest accuracy but also the small difference in confidence score between clear and low-light images, suggesting robust performance even in challenging lighting conditions.

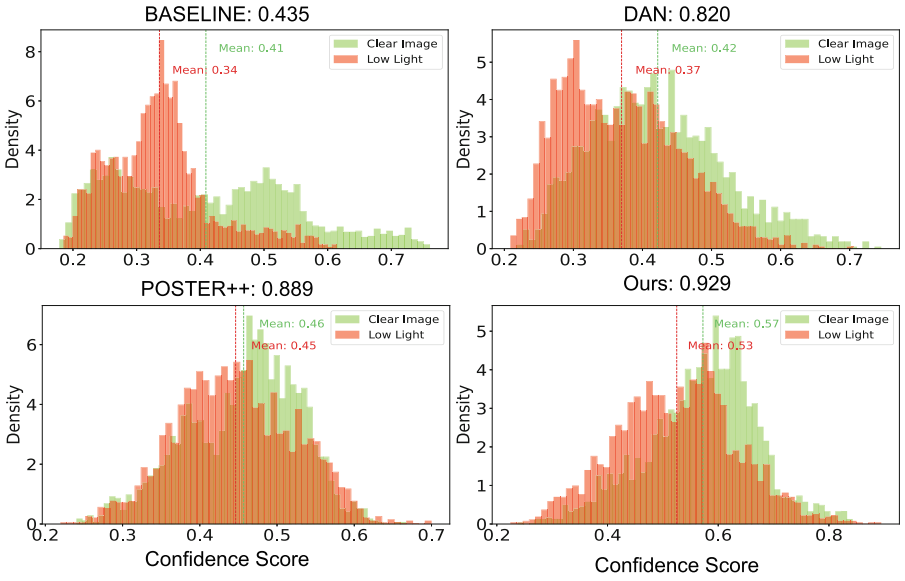


Fig. 6. Confidence score of different methods on KDEF dataset. Accuracy for each method is marked on the top. The baseline [16] method fails as FER data have small inter-class distances. DAN [19] and POSTER++ [9] have relative high confidence score while they still fall a lot in low confidence score area. Our method can effectively separates different emotion samples on clear and low-light images.

4.4 Ablation Study

This section evaluates the impacts of crucial components within LLDif, including the Diffusion Model (DM), various loss functions, and the insert noise during the training phase, as depicted in Table 5. (1) The contrast between LLDif$_{S2}$-V3 and LLDif$_{S2}$-V1 underscores the DM's robust ability in accurately predicting the embedding prior distribution EPD. (2) The insert noise into the DM's process in LLDif$_{S2}$-V4 is demonstrated to enhance the accuracy of EPD predictions. (3) The efficiency of different loss functions is also examined. The comparison between using $\mathcal{L}_{ce}$ in LLDif$_{S2}$-V4 (refer to Eq. (13)) and $\mathcal{L}_{total}$ in LLDif$_{S2}$-V2 (refer to Eq. (18)) shows that using $\mathcal{L}_{ce}$ is required for achieving better accuracy.

Impact of Iteration Numbers. This section examines how varying the number of iterations in the Diffusion Model (DM) influences the LLDif$_{S2}$ performance. We experimented with different iteration numbers in LLDif$_{S2}$, adjusting the β_t value (with α_t set as $1-\beta_t$, as outlined in Eq. 14) to ensure the variable Z evolves toward a Gaussian distribution, $Z_T \sim \mathcal{N}(0,1)$. Figures 8 and 7 demonstrate that LLDif$_{S2}$'s performance notably enhances at 4 iterations. Increasing the iteration number over 4 iterations does not substantially impact model's performance, suggesting the attainment of an optimal threshold. Notably, LLDif$_{S2}$

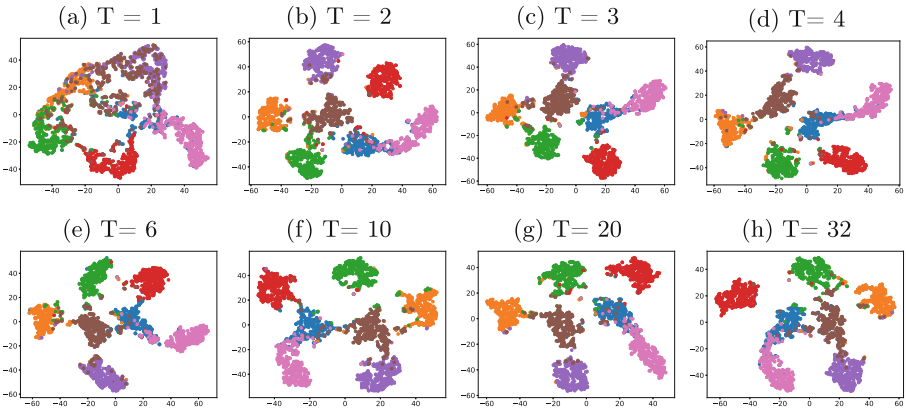


Fig. 7. Progressive clustering of features in diffusion space visualized using t-SNE at different time steps (T).

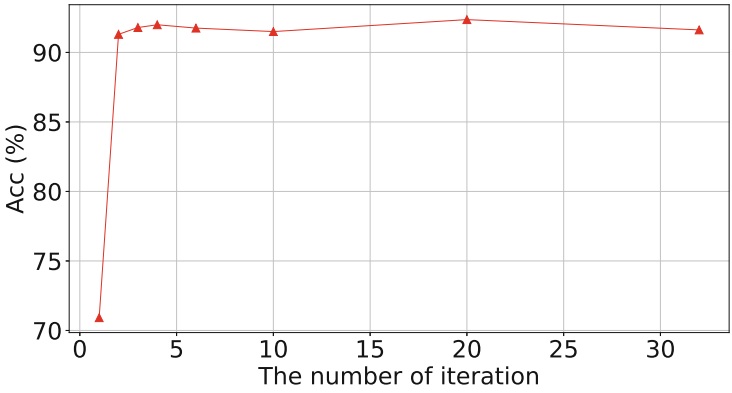


Fig. 8. Analyse impacts of iterations in DM.

reaches convergence more quickly than traditional DM methods, which typically requires over 50 iterations. This enhanced efficiency results from applying DM on the EPD, which is a one-dimensional, concise vector.

5 Conclusion

In this work, we present LLDif, an innovative framework utilising diffusion-based method to enhance facial expression recognition under low-light conditions. Addressing the challenges of image quality degradation in low-light settings, LLDif employs a two-stage training approach, utilizing a label-aware CLIP (LA-CLIP), an embedding prior distribution network (PNET), and a diffusion-based transformer network (LLformer). By integrating advanced architecture like the PNET and LLformer, LLDif can effectively restore emotion labels from degraded

low-light images at multiple scale. Our experiments confirms that LLDif outperforms existing methods, gains competitive performance on three low-light facial expression recognition datasets.

References

1. Barsoum, E., Zhang, C., Ferrer, C.C., Zhang, Z.: Training deep networks for facial expression recognition with crowd-sourced label distribution. In: ICMI (2016)
2. Farzaneh, A.H., Qi, X.: Facial expression recognition in the wild via deep attentive center loss. In: WACV (2021)
3. Han, X., Zheng, H., Zhou, M.: Card: classification and regression diffusion models. In: NeurIPS (2022)
4. Kollias, D., Zafeiriou, S.: Exploiting multi-CNN features in CNN-RNN based dimensional emotion recognition on the omg in-the-wild dataset. IEEE Trans. Affect. Comput. **12**(3), 595–606 (2020)
5. Kosugi, S., Yamasaki, T.: Unpaired image enhancement featuring reinforcement-learning-controlled image editing software. In: AAAI (2020)
6. Lee, I., Lee, E., Yoo, S.B.: Latent-ofer: detect, mask, and reconstruct with latent vectors for occluded facial expression recognition. In: CVPR (2023)
7. Li, S., Deng, W., Du, J.: Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild. In: CVPR (2017)
8. Liu, Y., Wang, W., Feng, C., Zhang, H., Chen, Z., Zhan, Y.: Expression snippet transformer for robust video-based facial expression recognition. Pattern Recogn. **138**, 109368 (2023)
9. Mao, J., Xu, R., Yin, X., Chang, Y., Nie, B., Huang, A.: Poster V2: a simpler and stronger facial expression recognition network. arXiv:2301.12149 (2023)
10. Niu, W., Zhang, K., Li, D., Luo, W.: Four-player GroupGAN for weak expression recognition via latent expression magnification. Knowl.-Based Syst. **251**, 109304 (2022)
11. Rahman, A., Valanarasu, J.M.J., Hacihaliloglu, I., Patel, V.M.: Ambiguous medical image segmentation using diffusion models. In: CVPR (2023)
12. Shang, S., et al.: Resdiff: combining CNN and diffusion model for image super-resolution. In: AAAI (2024)
13. She, J., Hu, Y., Shi, H., Wang, J., Shen, Q., Mei, T.: Dive into ambiguity: latent distribution mining and pairwise uncertainty estimation for facial expression recognition. In: CVPR (2021)
14. Shi, J., Zhu, S., Liang, Z.: Learning to amend facial expression representation via de-albino and affinity. arXiv preprint arXiv:2103.10189 (2021)
15. Vazquez-Rodriguez, J., Lefebvre, G., Cumin, J., Crowley, J.L.: Transformer-based self-supervised learning for emotion recognition. In: ICPR (2022)
16. Wang, K., Peng, X., Yang, J., Lu, S., Qiao, Y.: Suppressing uncertainties for large-scale facial expression recognition. In: CVPR (2020)
17. Wang, Z., Zhang, K., Luo, W., Sankaranarayana, R.: HTNet for micro-expression recognition. Neurocomputing **602**, 128196 (2024)
18. Wang, Z., Zhang, K., Sankaranarayana, R.: LRDif: diffusion models for under-display camera emotion recognition. arXiv preprint arXiv:2402.00250 (2024)
19. Wen, Z., Lin, W., Wang, T., Xu, G.: Distract your attention: multi-head cross attention network for facial expression recognition. Biomimetics **8**, 199 (2023)

20. Wu, J., Ji, W., Fu, H., Xu, M., Jin, Y., Xu, Y.: MedSegDiff-V2: diffusion-based medical image segmentation with transformer. In: AAAI (2024)
21. Yang, B., et al.: Paint by example: exemplar-based image editing with diffusion models. In: CVPR (2023)
22. Zhang, K., Huang, Y., Du, Y., Wang, L.: Facial expression recognition based on deep evolutional spatial-temporal networks. IEEE Trans. Image Process. **26**, 4193–4203 (2017)
23. Zhang, Q., et al.: Authentic emotion mapping: benchmarking facial expressions in real news. arXiv preprint arXiv:2404.13493 (2024)
24. Zhang, Q., Wang, Z., Liu, Y., Qin, Z., Zhang, K., Gedeon, T.: Geometric-aware facial landmark emotion recognition. In: 2023 6th International Conference on Software Engineering and Computer Science (CSECS) (2023)
25. Zhang, Y., Wang, C., Deng, W.: Relative uncertainty learning for facial expression recognition. In: NeurIPS (2021)
26. Zhang, Y., Wang, C., Ling, X., Deng, W.: Learn from all: erasing attention consistency for noisy label facial expression recognition. In: ECCV (2022)
27. Zhang, Z., Han, L., Ghosh, A., Metaxas, D.N., Ren, J.: Sine: single image editing with text-to-image diffusion models. In: CVPR (2023)
28. Zhao, Z., Liu, Q.: Former-DFER: dynamic facial expression recognition transformer. In: ACM MM (2021)
29. Zhao, Z., Liu, Q., Wang, S.: Learning deep global multi-scale and local attention features for facial expression recognition in the wild. IEEE Trans. Image Process. **30**, 6544–6556 (2021)

Dense Coordinate Channel Attention Network for Depression Level Estimation from Speech

Ziping Zhao[1(✉)], Shizhao Liu[1], Mingyue Niu[2], Haishuai Wang[3], and Björn W. Schuller[4,5]

[1] College of Computer and Information Engineering, Tianjin Normal University, Tianjin 300387, China
ztianjin@126.com, 2111090051@stu.tjnu.edu.cn
[2] School of Information Science and Engineering, Yanshan University, Qinhuangdao 066004, China
niumingyue@ysu.edu.cn
[3] College of Computer Science, Zhejiang University, Hangzhou 310058, China
haishuai.wang@zju.edu.cn
[4] Chair of Health Informatics, Technical University of Munich (TUM), Munich, Germany
[5] GLAM - Group on Language, Audio, and Music, Imperial College London, London, UK
bjoern.schuller@imperial.ac.uk

Abstract. Automatic depression level estimation from speech is currently an active research topic in the field of computational emotion recognition. One symptom commonly exhibited by patients with depression is erratic speech volume; thus, patients' voices can be used as a biosignature to identify their level of depression. However, speech signals have time-frequency properties; different frequencies and different timestamps contribute to depression detection in different ways. Accordingly, we design a Coordinate Channel Attention (CCA) block for differentiating tensor information with different contributions. We use a dense block to extract profound speech features with the above-mentioned blocks to form our proposed Dense Coordinate Channel Attention Network (DCCANet). Subsequently, a vectorization block is utilized to fuse the high-dimensional information. We split the original long speech into short audio segments of equal length, then feed these short segments into the network after feature extraction to determine BDI-II scores. Ultimately, the mean of the scores is used as the individual's depression level. Experiments on both the AVEC2013 and AVEC2014 datasets prove the effectiveness of DCCANet, which outperforms several existing methods.

Keywords: Coordinate Channel Attention · Depression Level Estimation · feature extraction · speech signals · time-frequency properties

A. Antonacopoulos et al. (Eds.): ICPR 2024, LNCS 15313, pp. 402–413, 2025.
https://doi.org/10.1007/978-3-031-78201-5_26

1 Introduction

Depression is a severe health disorder that can be distinguished from the usual mood swings and transient emotional reactions to the challenges of everyday life [18]. Traditional methods of diagnosing depression essentially count the number of presented symptoms; however, as depression becomes more severe, more concomitant symptoms can be observed [19], meaning that this method does not always support doctors in determining the patient's depression level or whether the patient is depressed. Moreover, the growing global burden of depression [12] indicates an urgent need for convenient and automated depression detection methods.

Table 1. BDI-II scores and corresponding levels of depression

BDI-II Score	Depression severity
0-13	None
14-19	Mild
20-28	Moderate
29-63	Severe

Physiological studies [3,13,14] have found significant differences in the speech signals of depressed and non-depressed patients. Accordingly, existing works that use deep learning techniques to find depression-related information in speech in order to determine depression levels are of significance. Psychometric evidence shows that the Beck Depression Inventory-II (BDI-II) [20,26] score can be utilized to estimate the severity of a person's depression, as shown in Table 1. Therefore, many studies [1,5,10,21,23,24] have used this score as a criterion for depression level assessment. Notably, although some of these methods have been proven to be effective to varying extents, some aspects of them could still be improved.

Some researchers [21,23,24] have used unsupervised hand-crafted features to estimate the severity of depression in individuals. However, this approach is highly dependent on the researcher's research orientation and professional experience; moreover, some depression-related information may be lost when using this approach. Li et al. [10] use GIE to obtain long-term global depression information and LASSO optimization to pool short-term features and thereby obtain long-term features. The TDCA network proposed by Cai et al. [1] uses dilated convolution blocks to extract time-domain speech signal information; this approach can aggregate multiscale contextual information associated with depression. However, while the above-mentioned methods have certain advantages, these authors focus only on the features of the time dimension in speech while neglecting the features of the frequency dimension. CSENet [5] has further demonstrated that the amplitude and phase spectrogram can be used to effectively perform feature extraction.

To deal with the above issue, we propose DCCANet to perform depression level estimation from speech. Specifically, we use the normalized amplitude and

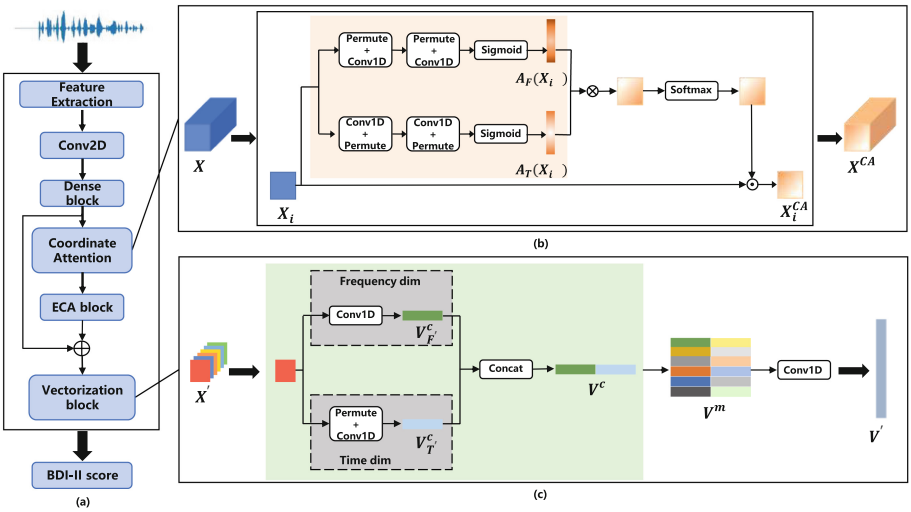


Fig. 1. (a) Detailed structure of the DCCA Network. (b) The coordinate attention (CA) block is used to highlight the time frames and frequency bands that make the largest contributions to each channel. (c) High-dimensional information fusion. *Permute* denotes matrix transposition

phase spectrogram as the input to the network, the Dense block to extract deep speech features, and the Coordinate Channel Attention (CCA) block to automatically select the time frames and frequency bands that make more significant contributions. Finally, inspired by [15], we utilize vectorization blocks to fuse high-dimensional information in order to differentiate between different levels of depression. Multiple experiments show that DCCANet is efficient for depression level estimation from speech. The main contributions of this paper can be summarized as follows:

- We propose a DCCANet for depression level estimation from speech. The CCA block can extract effective time frames, frequency bands, and channel information.
- We use the amplitude and phase spectrogram, normalized by time frame, as the input in order to help the network extract valid data.
- Experimental results on the AVEC2013 and AVEC2014 datasets prove that our method outperforms some of the more recent methods.

2 Methodology

Taguchi et al. [22] demonstrated that different time frames and frequency bands make different contributions to depression diagnosis. The attention mechanism can capture meaningful information from vocal inputs with high efficiency [7, 11, 27]. In light of the above, we propose DCCANet. Specifically, we use the

mean-variance normalized amplitude and phase spectrogram as the input of the model, extract the deep speech features via Conv2d and Dense block, then use our designed CCA block to filter the time frames and frequency bands that contribute most prominently, and finally fuse the high-dimensional information to predict the individual depression level. Our proposed framework is illustrated in Fig. 1(a).

2.1 Feature Extraction Module

To improve the extraction of speech information, we use the amplitude and phase spectrogram $X_p \in \mathrm{R}^{2\times \mathrm{F}\times \mathrm{T}}$, normalized by the time frame, as the input of the model, and extract the depth features of the speech using the dense block [8]. First, the short-time Fourier transform (STFT) is applied to the divided speech segments $x[s]$, as shown by Eq. (1):

$$\mathcal{A}[t,f] + \mathcal{P}[t,f] * i = \mathcal{STFT}(x[s]), \tag{1}$$

where $\mathcal{A}$ and $\mathcal{P}$ denote the corresponding amplitude and phase spectrograms, respectively. $\mathcal{STFT}$ denotes the short-time Fourier transform calculation. t, f, and s are the index of the time frames, the frequency bands, and the speech signals, respectively.

Subsequently, without affecting the physical meaning of the data, we normalize the resulting amplitude and phase spectrograms by the time frame, as shown in Eq. (2):

$$norm(x) = \frac{x_i - mean\,(x_i)}{var\,(x_i)}, \tag{2}$$

where x denotes the input speech segment, $norm(\cdot)$ denotes the normalization function, i denotes the i-th frame of x, $mean(\cdot)$ denotes the mean of the frame, and $var(\cdot)$ denotes the variance of the frame.

Finally, the normalized amplitude and phase parts are stacked together to form the input X_p to the model, as shown in Eq. (3):

$$X_p = stack(norm(\mathcal{A}[t,f]), norm(\mathcal{P}[t,f])), \tag{3}$$

where $stack(\cdot)$ denotes the tensor concatenation operation, X_p denotes the network framework input.

With the feature extraction module, we extract the depression information in the speech, while also taking into account the phase information contained in the Fourier complex spectrogram.

2.2 Coordinate Channel Attention Block

In this section, we present the details of the designed coordinate channel attention (CCA) block, which consists of a Coordinate Attention (CA) block and an Efficient Channel Attention (ECA) block. The CA block examines the contribution of each channel in the tensor for different time frames and frequency bands.

The ECA block helps the model to attend to essential channels and suppress noise or less informative channels, thereby improving the overall efficiency of the network. In our method, we use a Conv2D and Dense block to extract the high dimensional depression feature information of X_p in order to obtain the input $X \in \mathrm{R}^{\mathrm{C}\times\mathrm{F}\times\mathrm{T}}$ of CCA; here, T, F, and C denote the total count of timestamps (i.e., columns), the total count of frequency bands (i.e., rows), and the total count of channels, respectively.

Coordinate Attention. We define each channel $x_i \in \mathrm{R}^{\mathrm{F}\times\mathrm{T}}(i = 1, 2, \cdots, C)$ of X as an input to the CA block. Specifically, we compute the weight of the timestamps and frequency bands for x_i by Eqs. (4) and (5), respectively. Finally, the weighted channels that emphasize the depression information are obtained by Eq. (6).

$$A_T\left(x_i\right) = \sigma\left(con_T^2\left(con_T^1\left(x_i\right)\right)\right) \in R^{1\times T}, \tag{4}$$

$$A_F\left(x_i\right) = \sigma\left(con_F^2\left(con_F^1\left(x_i\right)\right)\right) \in R^{F\times 1}, \tag{5}$$

where con_F^1 and con_T^1 are two Conv1D operations along the time and frequency axes, respectively, while con_F^2 and con_T^2 are two Conv1D operations used to obtain the weights on the time and frequency axes. σ is the sigmoid activation function.

$$X_i^{CA} = X_i \odot \delta\left(A_F\left(X_i\right) \otimes A_T\left(X_i\right)\right), \tag{6}$$

where X_i^{CA} denotes the channel tensor after the weights are obtained, δ denotes the normalized exponential function, and $\odot$ and $\otimes$ denote the element-wise multiplication and the matrix multiplication operation, respectively.

By performing the above operation, we calculate the weights of each X_i to obtain $X^{CA} \in \mathrm{R}^{\mathrm{C}\times\mathrm{F}\times\mathrm{T}}$. Figure 1(b) illustrates the operation of the CA block.

Efficient Channel Attention Block. Wang et al. [25] proposed the ECA Network, which has proven very effective in a deep learning context. This module involves only a few parameters and provides significant performance gains without dimensionality reduction.

Specifically, the ECA block independently applies global average pooling to each channel, consolidating three-dimensional channel information into a one-dimensional vector. Moreover, a Conv1D with kernel size k is used to obtain the channel-level weights for the current channel and its k neighboring channels. Finally, the tensor is activated by a sigmoid function. The channel weight calculated by the ECA block is shown in Eq. (7):

$$X^{ECA} = \sigma\left(con_c(gap(X^{CA}))\right), \tag{7}$$

where $X^{E\ \mathrm{C}A} \in \mathrm{R}^{\mathrm{C}\times\mathrm{F}\times\mathrm{T}}$ is the output to the ECA block, σ is a sigmoid activation function, con_c denotes Conv1D, and gap denotes the use of global average pooling for each channel.

In summary, the Coordinate Channel Attention block calibrates all channels and examines the time frame and frequency band of each channel.

2.3 High-Dimensional Information Fusion

To obtain more accurate individual BDI-II scores, we utilize time-frequency channel vectorization block [15] to perform high-dimensional information fusion. In this block, two Conv1D operations are employed for each channel of input $\boldsymbol{X}' \in \mathrm{R}^{\mathrm{C}\times\mathrm{F}\times\mathrm{T}}$ to fuse the information for each time frame and frequency band in order to obtain $\boldsymbol{V}_{T'}^{c} \in \mathrm{R}^{1\times\mathrm{T}}$ and $\boldsymbol{V}_{F'}^{c} \in \mathrm{R}^{1\times\mathrm{F}}$, respectively. We then splice the two feature vectors to obtain tensor $\boldsymbol{V}^{\mathrm{c}} \in \mathrm{R}^{1\times(\mathrm{F}+\mathrm{T})}$, which contains depression information within a channel. Finally, according to Eq. (8), we concatenate each tensor into a new matrix in a row-by-row manner to fuse the depression representations of all channels.

$$V^{m} = \begin{bmatrix} v^{1'} \\ v^{2'} \\ \cdots \\ v^{c'} \end{bmatrix}, \tag{8}$$

where V^{m} is the vectorized representation of all channels and $v^{c'}$ is the vectorized representation of each channel.

Next, the fusion information of each channel is again fused by using Conv1D to get $\boldsymbol{V}' \in \mathrm{R}^{C}$. Figure 1(c) illustrates the high-dimensional information fusion process. Finally, we use a Fully Connected Layer to convert $\boldsymbol{V}'$ into BDI-II scores.

3 Experiments

3.1 Dataset and Evaluation Metrics

The AVEC2013 [24] dataset consists of 150 sample videos recording 84 subjects performing 14 different tasks based on computer prompts, with video lengths ranging from 20 min to 50 min. The AVEC2013 dataset is divided into three sections of equal size (each containing 50 samples): specifically, training, development, and test sets.

The AVEC2014 [23] dataset consists of two parts, "Northwind" and "FreeForm". The length of recordings for the Northwind task ranged from 31 s to 89 s, while the length of recordings for the Freeform task ranged from 6 s to 248 s. Each task has 150 videos, divided equally into three sections: training, development, and test sets. In our method, we merged the two subsets into a dataset. Thus, each set has 100 audio samples.

Both datasets used the BDI-II as a criterion for determining individual depression levels. The relevant studies apply mean absolute error (MAE) and root mean square error (RMSE) as assessment metrics for estimating the level of depression, which are shown in Eqs. (9) and (10):

$$MAE = \frac{1}{M}\sum_{i=1}^{M}\left|a_i - p_i\right|, \tag{9}$$

$$RMSE = \sqrt{\frac{1}{M}\sum_{i=1}^{M}\left(a_i - p_i\right)^2}, \tag{10}$$

where a and p denote the actual and predictive BDI-II scores, respectively, while i is the index of the sample.

3.2 Implementation Details

In the experiments conducted in this paper, we first divide the sample audio into segments of 3 s in length, with a repetition rate of 50% at the adjacent ends. The rate of sampling for all extracted speech signal waveforms is 8000 Hz. In Eq. (1), the Hamming window length is 32 ms and the window movement is 16 ms, meaning that each window contains 256 sampling points; thus, the size of the spectrograms is 129×188. The number of output channels for Conv2d is 64, the growth rate for the Dense block is 32, and the number of Dense blocks is 6; in short, the number of channels for profound speech features is 256. In the CCA block, we set the parameters to v_f=7; v_t=7; v_c=13, which represents the convolution kernel size in the frequency dimension, time dimension, and channel dimension, respectively. In this experiment, the optimizer utilized is Adam [9], the loss function used is RMSE, and the learning rate is 0.0001. We expanded the data to optimize the AVEC2014 model using AVEC2013's data and also optimized the AVEC2013 model using AVEC2014's data.

It is worth mentioning that in order to address the problem of the uneven number of speech samples at different levels of depression. We first put all the speech segments of the same subject into the same folder, and then divided all the subject files of the same depression level score into groups. When training the net-work, we took 8 speech segments sequentially from each group with the same depression level score and ensured that each subject segment was sampled evenly. With the above method, we ensure that the number of speech segments at each depression level is the same in each training epoch.

Table 2. Comparison of DCCANet with some existing systems on the AVEC2013 test set.

Systems	RMSE	MAE
Hybrid Net+lp-norm Pooling [16]	9.79	7.48
SAN-CNN [28]	9.65	7.38
LSTM sub-network with GIE [10]	9.63	7.51
STA+EEP [17]	9.5	7.14
CSENet [5]	9.28	6.79
TDCA-Net [1]	9.22	6.9
SR+SER [4]	8.73	7.32
CSF+GSR [2]	8.5	–
DCCANet(Ours)	**8.47**	**6.78**

Table 3. Comparison of DCCANet with some existing systems on the AVEC2014 test set.

Systems	RMSE	MAE
Hybrid Net+lp-norm Pooling [16]	9.66	8.02
SAN-CNN [28]	9.57	7.94
LSTM sub-network with GIE [10]	9.4	7.37
STA+EEP [17]	9.13	7.65
CSENet [5]	9.61	7.13
ADTP [6]	9.27	7.26
TDCA-Net [1]	8.9	7.08
SR+SER [4]	8.82	6.8
DCCANet(Ours)	**7.54**	**6.17**

3.3 Comparison with Other Methods

Tables 2 and 3 report our comparisons with existing methods on AVEC2013 and AVEC2014, respectively. Compared with the other baseline methods, especially neural networks that incorporate attention modules [1,5,10], our proposed model achieves better accuracy on both metrics. These findings indicate that our proposed DCCANet improves the detection of depression levels from speech. This is because the CCA block not only focuses on features in the channel and temporal dimensions but also highlights features in the frequency band dimension.

3.4 Effectiveness of the Feature Extraction Module

Figure 2 shows the feature extraction results of the subjects with different depression levels. Furthermore, to maintain a single variable, we selected the results of two subjects who performed the same task simultaneously. Here, the horizontal axis denotes the time dimension after feature extraction, while the vertical axis represents the frequency dimension after feature extraction. The brighter the color, the larger the value. As Fig. 2 shows, the features of healthy and depressed individuals differ significantly, especially the part enclosed by the black box; this indicates that the feature extraction block can effectively distinguish the differences in acoustic characteristics.

3.5 Effectiveness of the CCA Module

To further explore the contribution of DCCANet, we conducted ablation experiments using four network structures on two datasets (AVEC2013 and AVEC2014) to verify the contribution of each component in DCCANet. As shown by the experimental results (listed in Table 4), our proposed DCCANet improves depression level estimation from speech by calibrating the channels as well as by handling the time-frequency properties of each channel.

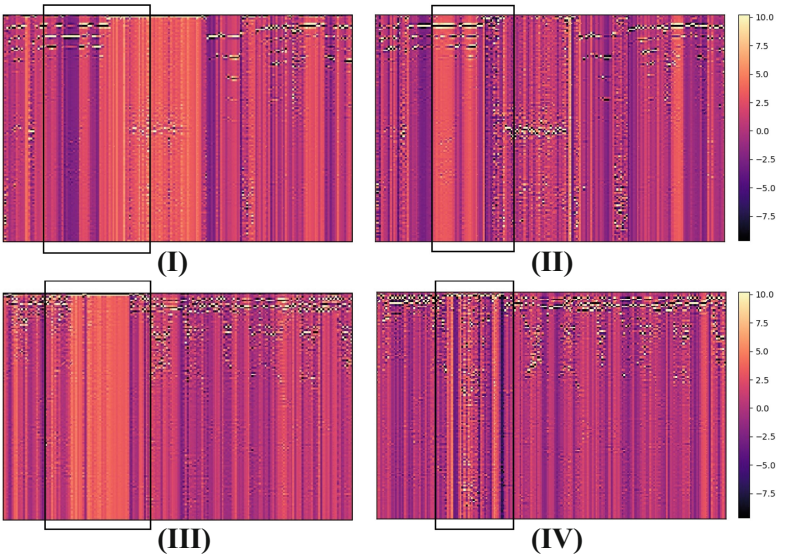


Fig. 2. (I) and (II) are the amplitude and phase spectrograms (respectively) obtained from feature extraction for healthy subjects No. 215-3. (III) and (IV) are the amplitude and phase spectrograms (respectively) obtained from feature extraction of No. 241-2 for major depressive disorder patients.

3.6 Visualization of Predictions

In order to visualize the prediction results of the method in this paper, we plotted a scatter plot as shown in Fig. 3. The plot shows the predicted values of the model and the true values of the subjects. The horizontal axis represents the subjects' IDs and the vertical axis represents the BDI-II scores. As the accuracy of the prediction of moderate depression may mask the inadequacy of the prediction of other depression levels, we tuned up the weights given to the loss values of healthy individuals and mild depression patients during training. As shown in

Table 4. Depression level estimation performance on the AVEC2013 and AVEC2014 test sets using various frameworks. "C_2" is an abbreviation for "Conv2D". "DB" is an abbreviation for "Dense Block".

Network structures	AVEC2013		AVEC2014	
	RMSE	MAE	RMSE	MAE
C_2+DB	9.12	7.15	8.32	6.93
C_2+DB+ECA	8.76	6.96	7.89	6.46
C_2+DB+CA	8.59	6.92	8.08	6.56
C_2+DB+CCA	**8.47**	**6.78**	**7.54**	**6.17**

the figure, the method in this paper has the best performance in predicting patients with moderate depression.

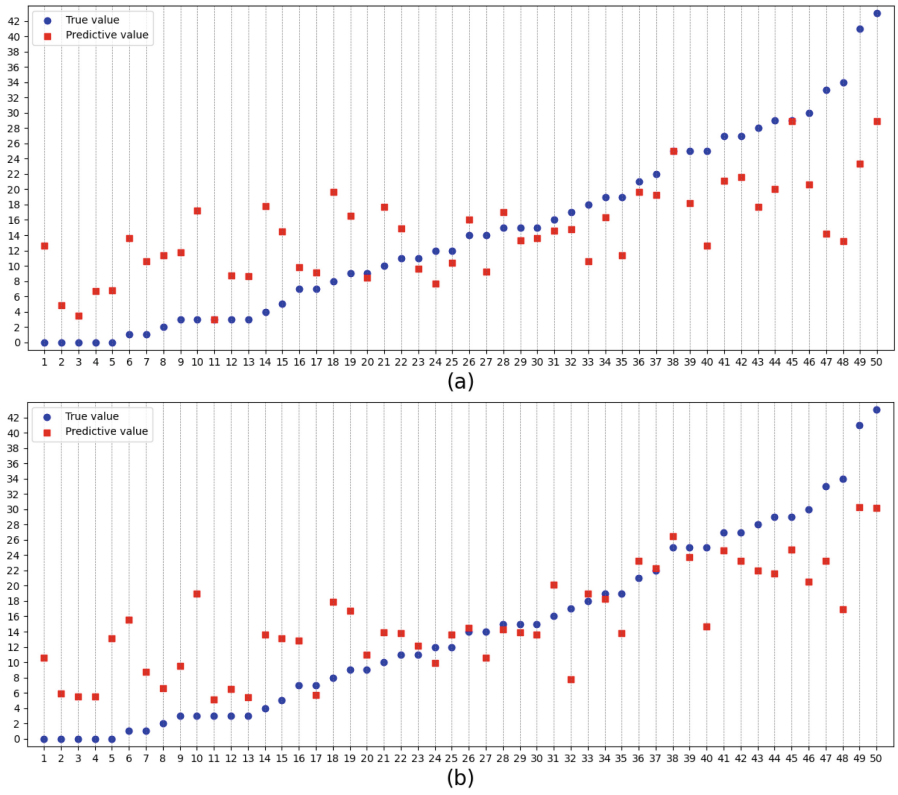


Fig. 3. The scatter plot of the true value vs predictive value on the test set of AVEC2013 (a) and AVEC2014 (b) based on DCCANet.

4 Conclusion

Different time frames and frequency bands make different contributions to predicting individual depression levels. Therefore, we proposed a DCCA network to estimate an individual's depression level by extracting high dimensional speech features of the normalized amplitude and phase spectrogram, then examining the time frame and frequency band of each channel, calibrating all channels, and finally fusing the high-dimensional features. Experiments and ablation studies on the AVEC2013 and AVEC2014 databases prove that the proposed CCA block can successfully differentiate temporal frames and frequency bands that make different degrees of contribution to depression features, and can also emphasize informative channels. On these two datasets, DCCANet obtains good performance compared to several previous existing methods.

In future work, we could investigate the embedding of these principles into other network architectures in order to find information about depression from speech more effectively.

Acknowledgements. The work is supported by the National Natural Science Foundation of China (No. 62071330) and the Open Project Program of the State Key Laboratory of Multimodal Artificial Intelligence System (No. 202200012).

References

1. Cai, C., Niu, M., Liu, B., Tao, J., Liu, X.: TDCA-Net: time-domain channel attention network for depression detection. In: Proceedings of the INTERSPEECH, pp. 2511–2515. Brno, Czechia (2021)
2. Cummins, N., Sethu, V., Epps, J., Williamson, J.R., Quatieri, T.F., Krajewski, J.: Generalized two-stage rank regression framework for depression score prediction from speech. IEEE Trans. Affect. Comput. **11**(2), 272–283 (2020)
3. Dietrich, M., Abbott, K.V., Gartner-Schmidt, J., Rosen, C.A.: The frequency of perceived stress, anxiety, and depression in patients with common pathologies affecting voice. J. Voice **22**(4), 472–488 (2008)
4. Dong, Y., Yang, X.: A hierarchical depression detection model based on vocal and emotional cues. Neurocomputing **441**, 279–290 (2021)
5. Fan, C., Lv, Z., Pei, S., Niu, M.: CSENet: complex squeeze-and-excitation network for speech depression level prediction. In: Proceedings of the 47th IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 546–550. Singapore (2022)
6. Fu, X., Li, J., Liu, H., Zhang, M., Xin, G.: Audio signal-based depression level prediction combining temporal and spectral features. In: Proceedings of the 26th International Conference on Pattern Recognition (ICPR), pp. 359–365. MontrÃal, Canada (2022)
7. Hu, J., Shen, L., Sun, G.: Squeeze-and-excitation networks. In: Proceedings of the 31st IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 7132–7141. Salt Lake City, Utah, USA (2018)
8. Huang, G., Liu, Z., Van Der Maaten, L., Weinberger, K.Q.: Densely connected convolutional networks. In: Proceedings of the 30th IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 4700–4708. Honolulu, Hawaii, USA (2017)
9. Kingma, D.P., Ba, J.: Adam: a method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2017)
10. Li, Y., Niu, M., Zhao, Z., Tao, J.: Automatic depression level assessment from speech by long-term global information embedding. In: Proceedings of the 47th IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 8507–8511. Singapore (2022)
11. Lin, S., Zeng, Y., Gong, Y.: Learning of time-frequency attention mechanism for automatic modulation recognition. IEEE Wireless Commun. Lett. **11**(4), 707–711 (2022)
12. Liu, Q., He, H., Yang, J., Feng, X., Zhao, F., Lyu, J.: Changes in the global burden of depression from 1990 to 2017: findings from the global burden of disease study. J. Psychiatr. Res. **126**, 134–140 (2020)

13. Mundt, J.C., Snyder, P.J., Cannizzaro, M.S., Chappie, K., Geralts, D.S.: Voice acoustic measures of depression severity and treatment response collected via interactive voice response (IVR) technology. J. Neurolinguistics **20**(1), 50–64 (2007)
14. Mundt, J.C., Vogel, A.P., Feltner, D.E., Lenderking, W.R.: Vocal acoustic biomarkers of depression severity and treatment response. Biol. Psychiat. **72**(7), 580–587 (2012)
15. Niu, M., Liu, B., Tao, J., Li, Q.: A time-frequency channel attention and vectorization network for automatic depression level prediction. Neurocomputing **450**, 208–218 (2021)
16. Niu, M., Tao, J., Liu, B., Fan, C.: Automatic depression level detection via LP-norm pooling. In: Proceedings of the INTERSPEECH, pp. 4559–4563. Graz, Austria (2019)
17. Niu, M., Tao, J., Liu, B., Huang, J., Lian, Z.: Multimodal spatiotemporal representation for automatic depression level detection. IEEE Trans. Affect. Comput. **14**(1), 294–307 (2020)
18. Organization, W.H., et al.: Depression and other common mental disorders: global health estimates. Tech. rep., World Health Organization (2017)
19. Paykel, E.S.: Basic concepts of depression. Dialogues Clin. Neurosci. **10**(3), 279–289 (2022)
20. Skaik, R.S., Inkpen, D.: Predicting depression in Canada by automatic filling of beck's depression inventory questionnaire. IEEE Access **10**, 102033–102047 (2022)
21. Stasak, B., Epps, J., Goecke, R.: An investigation of linguistic stress and articulatory vowel characteristics for automatic depression classification. Comput. Speech Lang. **53**, 140–155 (2019)
22. Takaya, T., et al.: Major depressive disorder discrimination using vocal acoustic features. J. Affect. Disord. **225**, 214–220 (2018)
23. Valstar, M., et al.: AVEC 2014: 3D dimensional affect and depression recognition challenge. In: Proceedings of the 4th International Workshop on Audio/Visual Emotion Challenge (AVEC), pp. 3–10. Orlando, Florida, USA (2014)
24. Valstar, M., et al.: AVEC 2013: the continuous audio/visual emotion and depression recognition challenge. In: Proceedings of the 3rd ACM International Workshop on Audio/Visual Emotion Challenge (AVEC), pp. 3–10. Barcelona, Spain (2013)
25. Wang, Q., Wu, B., Zhu, P., Li, P., Zuo, W., Hu, Q.: ECA-Net: efficient channel attention for deep convolutional neural networks. In: Proceedings of the 33rd IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 11534–11542. Seattle, USA (2020)
26. Wang, Y., Gorenstein, C.: Psychometric properties of the beck depression inventory-ii: a comprehensive review. Braz. J. Psychiatry **35**(4), 416–431 (2013)
27. Wang, Y., et al.: Transformer-based acoustic modeling for hybrid speech recognition. In: Proceedings of the 45th IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), pp. 6874–6878. Virtual, Barcelona (2020)
28. Zhao, Z., et al.: Hybrid network feature extraction for depression assessment from speech. In: Proceedings of the INTERSPEECH, pp. 4956–4960. Shanghai, China (2020)

Interpreting Emotions Through the Grad-CAM Lens: Insights and Implications in CNN-Based Facial Emotion Recognition

Jens Gebele[1,2(✉)], Philipp Brune[1], Frank Schwab[2], and Sebastian von Mammen[2]

[1] Neu-Ulm University of Applied Sciences, Wileystraße 1, 89231 Neu-Ulm, Germany
jens.gebele@hnu.de
[2] University of Würzburg, Am Hubland, 97074 Würzburg, Germany

Abstract. Facial Emotion Recognition (FER) has gained significant attention in recent years due to its potential applications in various fields such as automotive, mental health, and education. Despite the impressive results of Deep Learning (DL) in these areas, a critical shortfall of these systems is the lack of explainability. This paper presents a systematic analysis using Gradient-weighted Class Activation Mapping (Grad-CAM) combined with Guided Backpropagation to investigate the features learned by DL models in FER and their alignment with established emotional theories. We apply this methodology to Convolutional Neural Networks (CNNs) trained on three different emotional datasets: FER-2013, RAF-DB, and AffectNet. Our findings indicate that machine-learned features vary within an emotional state and are not necessarily aligned with expert understanding in emotion psychology. This raises questions about the reliability and ethical implications of using FER systems in sensitive areas, where accurate interpretation of emotions is critical. In response, our study proposes exploring Neuro-Symbolic AI approaches as a potential pathway to more effectively grasp the complexity of emotion psychology and address these concerns. This approach paves the way for the development of new FER model architectures, potentially fostering the emergence of more nuanced emotional concepts.

Keywords: Grad-CAM · Guided Grad-CAM · Explainable AI · Emotion Recognition · Facial Emotion Recognition · Convolutional Neural Network

1 Introduction

Emotion recognition refers to the capability of technical systems to infer human emotional states from various behavioral cues, utilizing different data modalities. Among these, facial emotion recognition has gained significant popularity due

A. Antonacopoulos et al. (Eds.): ICPR 2024, LNCS 15313, pp. 414–429, 2025.
https://doi.org/10.1007/978-3-031-78201-5_27

to the strong link between facial expressions and the external manifestation of inner emotional states [19]. The concept of emotions is underpinned by numerous theories that describe different facets of emotional experiences. The foundational framework for many research activities in this field is based on Paul Ekman's proposal of seven universal emotional states: surprise, anger, disgust, fear, happiness, sadness, and contempt [9]. A neutral expression can be added to these. Due to the apparent simplicity of employing discrete emotional categories in Machine Learning (ML) applications, these labels constitute the foundation for many (research) datasets in the field. High-quality annotations typically require substantial efforts and are best performed by certified experts knowledgeable in emotional theory [21]. However, subjectivity is inherent to any labeling process, irrespective of the annotators' level of expertise. Furthermore, there has been a lively debate in the emotion research community regarding the universal applicability or validity of Ekman's theory [4]. This discussion is particularly relevant in light of earlier work indicating potential inconsistencies in data annotations [12], which may contribute to the challenges associated with both the subjective nature of emotional labeling and the ongoing debate about the universality of Ekman's framework.

In this context, it becomes crucial to examine which features ML models, particularly DL techniques such as CNNs, actually learn from limited or potentially inconsistent annotated facial emotion data. Despite their success in computer vision tasks, CNNs suffer from a lack of interpretability, which is particularly problematic in sensitive areas, such as FER, where transparency and comprehensibility are essential for ensuring reliable system operation [5]. Explainable AI (XAI) offers valuable techniques to investigate, after training (ex post), why CNN models make specific predictions. These techniques provide insights into the features or facial regions that contribute to the decision-making process and shed light on how these regions align with theories of emotion psychology [1].

The present study is organized as follows: Sect. 2 delves into attribution-based XAI techniques, and XAI in the context of FER. In Sect. 3, we outline the research methodology and describe the data used for our systematic analysis. The results obtained from applying Grad-CAM combined with Guided Backpropagation to explore learned features and facial regions within one CNN architecture are presented in Sect. 4. This is followed by an in-depth discussion in Sect. 5, which examines the implications of these findings with respect to emotion psychology theories. In addition, we offer insights into how future research in the realm of FER (and XAI) can benefit from Neuro-Symbolic AI approaches. Finally, we summarize our key contributions.

2 Related Work

XAI encompasses a range of techniques aimed at making AI decision-making transparent and interpretable. Within this broad domain, attribution-based methods are particularly significant in the field of computer vision.

2.1 Attribution-Based XAI Methods

Attribution-based methods, also referred to as pixel attribution, identify pixels that significantly contribute to the classification of an individual image by an Artificial Neural Network (ANN). Attribution-based methods focuses solely on image data [1,26], and can therefore be considered a subcategory of feature attributions. They allow to understand individual predictions of DL models by assigning a positive or negative value to each input feature proportional to its influence on the prediction task. Attribution-based methods are divided into two main categories. The first encompasses perturbation-based methods (also known as occlusion-based methods), such as SHAP (SHapley Additive exPlanations) [22] and LIME (Local Interpretable Model-agnostic Explanations) [29]. These techniques are model-agnostic and they generate explanations by modifying or perturbing the input in various ways to observe the impact on the model's prediction. The second category consists of methods that utilize the gradients of an ANN. Specifically, these methods compute the gradients of a particular classification output with respect to the input image. Depending on the approach, the way how gradients are calculated differ [1,26].

Several gradient-based XAI techniques have been developed to provide insights into how CNNs make predictions. In the following, we describe some prominent methods, which have evolved over time: Vanilla Gradient calculates the gradient of the output concerning the input image, highlighting pixels that significantly impact the model's decision. The resulting Saliency Map offers a visual representation of this influence [33]. DeconvNet aims to identify features that trigger certain layers in CNNs, i.e. DeconvNet methodologically reverses the operations of a CNN. This process projects feature activations back into the input pixel space, facilitating visualization of the relevant characteristics [37]. Class Activation Mapping (CAM) generates heatmaps which display class activation over input images, utilizing the final convolutional layer's weights in CNNs. By emphasizing critical regions, CAM highlights the important areas for prediction classes [38]. Grad-CAM unites gradient information with class activation maps to pinpoint image regions crucial for specific predictions. This technique yields more fine-grained heatmaps than CAM, enhancing localization capabilities [30]. Guided Grad-CAM integrates Guided Backpropagation and Grad-CAM techniques. Guided Grad-CAM produces high-resolution visualizations. These heatmaps emphasize not only significant regions but also the intricate details within those regions, which helps to understand the prediction process [30]. Smooth Grad improves the vanilla gradient method by averaging the gradients of the input image with added noise multiple times. This technique produces smoother and more visually coherent attribution maps, effectively reducing noise and highlighting salient features more clearly [34].

2.2 XAI in Facial Emotion Recognition

The major challenge in FER is to infer emotional (inner) states based on the outward expressions in the form of facial expressions, more generally referred

to as facial behavior. The majority of research on FER concentrates on FER model performance including the comparison of different FER model architectures (VGG-16, ResNet152, Inception, and Xception) and the ambition to achieve higher accuracy [11,17,28]. So far, XAI research on FER has witnessed limited exploration. Mouakher et al. [28] presented a multi-criteria evaluation framework for FER with various criteria such as accuracy and explainability. Shahabinejad et al. [31] also published a framework which is based on hybrid features, combining feature maps of face recognition with FER. Deramgozin et al. [8] discussed a hybrid XAI framework for FER, which consists of a CNN model and integrates explainable components based on LIME and facial Action Units. Guerdan et al. [15] work shows how facial affect analysis can contribute to make human-machine interactions more understandable. Shingjergji et al. [32] present an approach of gamified data collection of facial expression that contributes to explainability in FER.

Having established the broader context of XAI in FER, we now narrow our focus to Grad-CAM in FER. Bai et al. [3] discuss the analysis of mechanisms of spatio-temporal models for micro-expression recognition with Grad-CAM. Araf et al. [2] use Grad-CAM visualizations on a less complex real-time FER Cascade Classifier model than CNNs and based on only one dataset. Chen et al. [6] use Grad-CAM in combination with the so-called Broad Learning System for FER to explore the effects of structural system changes. Wadhawan and Gandhi [35] apply Grad-CAM to analyse Transfer Learning of Facial Landmark Localization within an ensemble network for FER. Malek-Podjaski and Deligianni [24] use a multi-encoder-decoder architecture to differentiate between biometrics and affects in human-motion affect recognition. Thereby, Grad-CAM is utilized to analyse the model in more depth. A very recent work [27] uses Grad-CAM to analyze the classification of facial AUs based on one emotional dataset.

3 Methodology

This study extends the data collection, preprocessing, and CNN model training methodologies developed in our previous work [12] where we evaluated the performance (precision, recall, and F1-score) of CNN models for seven universal emotional states across three datasets and identified inconsistencies in data annotations. Building on this foundation, the current work conducts a systematic analysis using slightly modified training settings and the same CNN model architectures. We employ attribution-based XAI methods to explore in greater depth how these models interpret emotional expressions, employing techniques analogous to [12].

3.1 Facial Emotion Databases

In our previous research [12], we conducted a comprehensive review of over 40 facial expression databases. We specifically focused on static databases that included annotations of the six basic emotions according to Ekman [9], plus

a neutral expression. From this selection, we excluded databases that featured multimodal and three-dimensional data, centering our attention solely on two-dimensional, static facial datasets. The remaining databases varied in terms of size, type of data (static or sequential), and the environment in which the data was collected (controlled or uncontrolled). We further narrowed down the selection by excluding datasets with fewer than 10,000 instances or those collected in controlled settings, aiming to work with a large, diverse, and more representative dataset.

Consequently, we focused on three primary databases. The first was the FER-2013 dataset [14], comprising 35,887 grayscale images. These images were automatically cropped and labeled, and validated by experts. The dataset includes seven emotional categories, with each image resized to 48×48 pixels [14]. The second database was the RAF-DB with basic emotions [20], which includes a total of 15,339 aligned, RGB-color images. These images were manually annotated by approximately 40 experts and resized to 100×100 pixels [20]. Lastly, the AffectNet (Mini Version) [25] was utilized. This subset includes only manually annotated RGB-color images, totaling 291,650 images, each with dimensions of 224×224 pixels. We chose to exclude the emotional state of contempt from our analysis to align with the seven emotions presented in the other datasets.

3.2 Data Pre-processing

Since the images were already aligned and cropped, our pre-processing step was confined to data normalization and augmentation, bypassing the need for face localization and facial landmark identification. Furthermore, we standardized the image sizes of RAF-DB and AffectNet to match the 48px * 48px resolution of FER-2013, ensuring uniformity for comparative analysis. Our normalization process involved scaling the RGB color values, which are represented by byte values, to a range of 0 to 1 by dividing by 255. The distribution of emotional classes across all datasets is detailed in Table 1.

Table 1. Distribution of Emotional Classes per Dataset

Emotion	FER-2013	RAF-DB	AffectNet
Angry	4,953	867	25,382
Disgust	547	877	4,303
Fear	5,121	355	6,878
Happy	8,989	5,957	134,915
Sad	6,077	2,460	25,959
Surprise	4,002	1,619	14,590
Neutral	6,198	3,204	75,374
Total	35,887	15,339	287,401

Most emotional states are well-represented across the three datasets, with a few exceptions. We partitioned each dataset into training and test sets in a 80:20 ratio, designating 30% of the training set for validation. This allocation resulted in 56% of the data for training, 24% for validation, and 20% for testing. These splits were stratified to maintain proportional representation of each emotional class in the training, validation, and test sets. Due to the smaller test set size in AffectNet, we combined its training and test sets before re-dividing them to align with the training, validation, and test set ratios of the other datasets.

3.3 CNN Model Training

Our CNN model, inspired by the AlexNet architecture [18] with its stacked convolutional layers, was designed not to exceed predefined performance benchmarks, but to enable a systematic comparison of Grad-CAM visualizations. It features four repeated units, each comprising two convolutional layers followed by max pooling. All convolutional layers use 'same' padding and ReLU activation. This architecture then transitions through a flattening step to two dense layers with dropout, ultimately classifying seven emotional states. The detailed specifications of our CNN model are provided in Table 2. To guarantee reproducibility and to minimize training variability, we standardized seed parameters and trained five models per dataset. Each model underwent 50 epochs of training with a batch size of 128 to ensure consistent weight updates, though the number of steps per epoch varied with dataset size. We utilized the Adam Optimizer at a constant learning rate of $1e-4$. The final model selected was the one achieving the highest validation accuracy across these epochs.

Table 2. Architecture of the CNN Model

	Layer	Output Shape	Parameters
4x{	2x Conv2D	(None, 48, 48, 32)	2,432 & 25,632
	MaxPooling2D	(None, 24, 24, 32)	0
	...		
	Flatten	(None, 2,304)	0
	Dense	(None, 128)	295,040
	Dropout	(None, 128)	0
	Dense	(None, 64)	8,256
	Dropout	(None, 64)	0
	Dense	(None, 7)	455

3.4 Grad-CAM and Guided Grad-CAM

Grad-CAM combines gradient data and class activation maps to identify critical image regions for specific predictions. This approach produces more detailed

heatmaps compared to CAM, thereby improving localization capabilities [30]. Previous research indicates that deeper layers of CNNs are responsible for learning more abstract, higher-level visual features [23]. This insight supports our rationale for focusing on the last convolutional layer, which offers an optimal balance between high-level feature representation and detailed spatial content. The referenced work underscores that as layers progress, they transition from capturing low-level to high-level features, making the final layers particularly effective for detailed yet abstract visual analysis. Applying Grad-CAM results in a class discriminative localization map L_c Grad-CAM $\in \mathbb{R}^{u \times v}$ represented in Eq. 1 [30].

$$L^c_{\text{Grad-CAM}} = \text{ReLU}(\sum_k w^c_k A^k_{ij})) \tag{1}$$

To this end, we propagate the image data through the CNN and determine the output score for the class of interest, right before the softmax layer. The output score of the other classes are ignored by setting their activations to zero. Next, the gradient of the class of interest is propagated backwards to the last convolutional layer. This operation is noted as Eq. (2) where y_c is the score of the class and A_k is the activation of each feature map k. The width and height dimensions are represented by the indices i and j, respectively. Subsequently, each 'pixel' within every feature map is assigned a weight based on its gradient with respect to the target class, as determined by the last convolutional layer [30].

$$\frac{dy^c}{dA^k_{ij}} \tag{2}$$

Then, Global Average Pooling is applied to the backpropagated gradients considering width i and height j. This results in weight w^c_k denoted as Eq. 3. w^c_k acts as a simplified representation of the deeper portions of the network, and quantifies the 'importance' of the feature map k in determining the output for a specific target class c [30].

$$w^c_k = \underbrace{\frac{1}{Z}\sum_i \sum_j}_{\text{Global Average Pooling}} \overbrace{\frac{dy^c}{dA^k_{ij}}}^{\text{Backpropagated Gradients}} \tag{3}$$

We proceed by multiplying each pixel's value of the feature map A_k with the corresponding gradient in order to receive the gradient-weighted average. Then the Rectified Linear Unit (ReLU) function is applied on the results to normalize them in range between 0 to 1. This is denoted as Eq. 4. By using ReLU, we focus on the information in the sense of features which have a positive contribution to the predicted class. The size of the heatmap corresponds to the size of the last convolutional layer feature maps (in our case 6×6). In a final step, we increase the resolution of the heatmaps to the original resolution of the image

size with 48×48 by means of bilinear interpolation [30].

$$L^c_{\text{Grad-CAM}} = \text{ReLU}\Big(\underbrace{\sum_k w^c_k}_{\text{Class Feature Weights}} \underbrace{A^k_{ij}}_{\text{Feature Map}} \Big) \tag{4}$$

Guided Grad-CAM merges the class-discriminative capabilities of Grad-CAM with the detailed, pixel-level visualization from Guided Backpropagation. While Grad-CAM outlines general regions of interest via coarse heatmaps, it may not clarify why certain features influence classifications. To enhance clarity, Guided Grad-CAM fuses these methods by element-wise multiplication of their outputs, after upscaling the Grad-CAM map to match the input image's resolution. This technique produces detailed visualizations that highlight precise features significant to the classification while maintaining an overall focus on relevant areas [30].

In our analysis, we generated visualization of Grad-CAM combined with Guided Backpropagation using the last convolutional layer of the best among five trained models for each of the seven emotional states. These were applied across three datasets: FER-2013, RAF-DB, and AffectNet. The visualizations from the leading model, trained on RAF-DB, formed the basis of our detailed analysis. Lastly, we computed aggregated Grad-CAM visualizations - capturing both means and standard deviations - for each emotional state and each model individually, as well as for all models collectively, across these datasets. This approach enabled us to explore potential variances among the models attributable to training differences.

4 Results

In this section, we first present the classification performance of our CNN model across three datasets: FER2013, RAF-DB, and AffectNet. Table 3 displays the average precision, recall, and F1-score from five training runs of the CNN model. We also provide selected Grad-CAM visualizations based on the RAF-DB dataset, as detailed in Sect. 3, focusing on the emotional states of 'disgust', 'sadness', and 'happiness'. These states were chosen based on their distinctive accuracy profiles across all datasets, with 'disgust' showing notably low accuracy, 'sadness' serving as the semantic opposite of 'happiness', and 'happiness' demonstrating consistently high accuracy.

Figure 1 shows four Grad-CAM visualizations. Three are aligned with the emotions of 'disgust' (Fig. 1a), 'sadness' (Fig. 1b), and 'happiness' (Fig. 1c), while Fig. 1d illustrates only false predictions within the 'happiness' category, contrasting with Fig. 1c, which contains only accurate predictions for 'happiness'. To aid in identification, red crosses mark all incorrectly classified images within these figures. This differentiation allows us to explore the discrepancies in activation areas between correctly and incorrectly predicted emotions.

The visualizations were generated using the Grad-CAM method combined with Guided Backpropagation, applied to the best-performing CNN model using

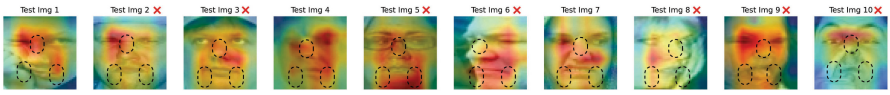
the last convolutional layer. We analyzed ten randomly selected images from the RAF-DB dataset's test set. Additionally, we manually highlighted key facial regions with dashed, circular markings, drawing from emotion psychology theories discussed in depth in Sect. 5.

Table 3. Average Precision, Recall, and F1-Score in five Model Training runs.

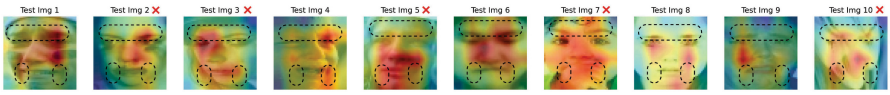
Emotion	Precision	Recall	F1-Score	Count
Angry-FER	0.49(±0.01)	0.46(±0.02)	0.48(±0.01)	991
Disgust-FER	0.83(±0.06)	0.16(±0.05)	0.26(±0.08)	109
Fear-FER	0.41(±0.01)	0.35(±0.04)	0.38(±0.02)	1,024
Happy-FER	0.77(±0.01)	0.77(±0.02)	0.77(±0.00)	1,798
Sad-FER	0.42(±0.02)	0.48(±0.04)	0.45(±0.01)	1,216
Surprise-FER	0.72(±0.05)	0.68(±0.03)	0.70(±0.01)	800
Neutral-FER	0.49(±0.03)	0.56(±0.04)	0.52(±0.01)	1,240
Angry-RAF	0.60(±0.05)	0.57(±0.04)	0.58(±0.03)	173
Disgust-RAF	0.44(±0.06)	0.26(±0.03)	0.33(±0.03)	175
Fear-RAF	0.67(±0.14)	0.32(±0.01)	0.43(±0.03)	71
Happy-RAF	0.84(±0.01)	0.88(±0.02)	0.86(±0.00)	1,192
Sad-RAF	0.61(±0.03)	0.57(±0.05)	0.59(±0.01)	492
Surprise-RAF	0.70(±0.04)	0.66(±0.04)	0.68(±0.01)	324
Neutral-RAF	0.61(±0.01)	0.72(±0.04)	0.66(±0.01)	641
Angry-Aff	0.41(±0.05)	0.31(±0.06)	0.35(±0.02)	5,076
Disgust-Aff	0.02(±0.04)	0.00(±0.00)	0.00(±0.00)	861
Fear-Aff	0.19(±0.17)	0.04(±0.04)	0.06(±0.06)	1,376
Happy-Aff	0.75(±0.01)	0.90(±0.01)	0.82(±0.00)	26,983
Sad-Aff	0.52(±0.05)	0.07(±0.04)	0.12(±0.07)	5,192
Surprise-Aff	0.31(±0.04)	0.22(±0.06)	0.25(±0.05)	2,918
Neutral-Aff	0.53(±0.01)	0.63(±0.04)	0.57(±0.01)	15,075

The visualizations consistently show varying activations within facial regions for each of the three emotional states. Notably, such variations are apparent across all seven emotional states and for all the three datasets - FER-2013, RAF-DB, and AffectNet.

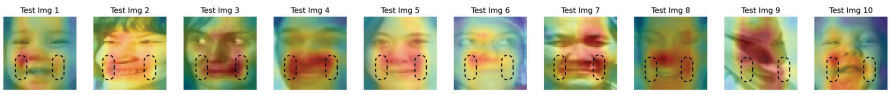
Alongside individual Grad-CAM visualizations, we employed aggregated Grad-CAM visualizations that capture both means and standard deviations across each emotional state to examine potential randomness in CNN model training. These visualizations revealed variability in activations among the five trained models, attributable to the inherent randomness in the training process. Despite this variability, a consistent finding across all three datasets is the fluctuation of specific facial features activated within any given emotional class.



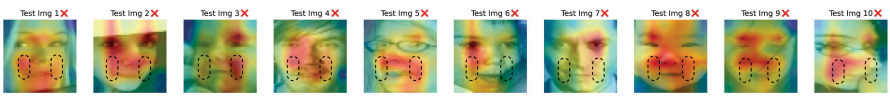
(a) Visualizations for 'Disgust' using the Final Convolutional Layer of the CNN Model.



(b) Visualizations for 'Sadness' using the Final Convolutional Layer of the CNN Model.



(c) Visualizations for Correctly Classified Images of the 'Happiness' class using the Final Convolutional Layer of the CNN Model.



(d) Visualizations for Incorrectly Classified Images of the 'Happiness' class using the Final Convolutional Layer of the CNN Model.

Fig. 1. Visualizations of Grad-CAM combined with Guided Backpropagation for 'Disgust', 'Sadness', and 'Happiness' using images from the RAF-DB dataset.

Thus, while activations may differ between the training of five CNN models, the features activated within the same emotional class also exhibit variability within each individual model.

5 Discussion

Our systematic analysis of CNN models using Grad-CAM combined with Guided Backpropagation reveals insightful nuances on how these models process and interpret facial expressions of emotion. While these CNN models achieve satisfactory performance scores, especially with emotions like 'happiness', the primary goal was to understand the knowledge they encode. We observe notable inconsistencies in feature activation and their intensities within the same emotional class across various datasets.

To contextualize these findings, comparing the learned features for individual emotions against established emotion psychology expertise is instructive. Paul Ekman's Facial Action Coding System (FACS) describes facial activities based on Action Units (AUs) [10]. According to FACS and related research [7,13,36], an emotion prediction table describes the relationship between AU combina-

tions and emotional states. Table 4 summarizes the intersections of AU-emotion combinations derived from multiple studies [7,13,36].

Table 4. AUs associated with Different Emotions.

Emotion	AUs
Happy	6 +12
Sadness	1 + 4 + 15
Disgust	9 + 15
Anger	4 + 5 + 7 + 23
Fear	1 + 2 + 4 + 5 + 20
Surprise	1 + 2 + 5 + 26

To accurately recognize 'Disgust', activation of both AU 9, the 'Nose Wrinkler', and AU 15, the 'Lip Corner Depressor', is essential. AU 9 features a muscle that stretches from the root of the nose to a point next to the nostril wings, contracting to lift the skin below the nostril wings towards the nose's root. AU 15, which originates from the side of the chin and attaches near the lip corner, draws the corners of the lips downward [10].

Figure 1a shows ten visualizations of 'Disgust', which is the emotional class with the lowest accuracy for all three datasets. In these visualizations the regions associated with AU 9 and 15 are highlighted with dotted, circular elements. Analysis of truly predicted images such as Test Image 1, 4, and 7 reveals consistent activations in the AU 9 area, whereas activations in the AU 15 region are either absent, weak, or unilateral. Conversely, in falsely predicted images like Test Image 2, 5, 9, and 10, a similar trend can be observed, although Test Image 5 notably exhibits perfect alignment of activations with AU 9 and 15.

For the emotional state 'sadness', expected activations include AU 1, the 'Inner Brow Raiser', AU 4, the 'Brow Lowerer', and again AU 15 the 'Lip Corner Depressor'. AU 1 involves a large muscle in the scalp and forehead that raises the eyebrows, running vertically from the top of the head to the eyebrows and covering almost the entire forehead. AU 4 involves three muscle strands in the forehead that act together to modify eyebrow position [10].

Analyzing Fig. 1b reveals that in correctly classified images, regions associated with AU 1 and AU 4 are only moderately activated in Test Images 1 and 4. Conversely, these activations are completely absent in Test Images 6, 8, and 9. Notably, incorrectly classified images tend to show stronger activations in the AU 1 and AU 4 areas. Similarly, for AU 15, correctly classified images exhibit less intense activations compared to those that are incorrectly predicted.

For accurate recognition of 'happiness', both AU 12, the 'Lip Corner Puller', and AU 6, the 'Cheek Raiser' must be present. The muscle for AU 12 is positioned high in the lower face, near the cheekbones, and stretches to the corners of the lips. It acts to pull the corners of the lips upward toward the cheekbones at an

oblique angle. In contrast, the muscle associated with AU 6 is located in the lower face, extending from the cheekbones to the corners of the eyes [10]. The regions affected by theses AUs are highlighted in Fig. 1c and 1d.

Figure 1c demonstrates that, for some images in the test set, the activated regions align perfectly with AU 6 and AU 12, as seen in Test Image 3, 4, and 5. However, other cases, such as Test Image 1 and 6, activation is only unilateral and moderate. Additionally, Test Image 7 and 9 show (large) activation in regions outside of AU 6 and AU 12, which do not correspond well with the expected regions. A comparison between Fig. 1c, which contains only true predictions, and Fig. 1d, which includes solely false predictions, reveals a generally similar activation pattern. However, Test Image 6 and 7 of Fig. 1d are notable exceptions, where incorrect activations are specifically centered around the left and right eye, respectively.

Our analysis of the emotional expressions of 'disgust', 'sadness', and 'happiness', reveals discrepancies between CNN-based feature activations and traditional interpretations of emotion psychology. For 'disgust', we observed little variation between activations in correctly and incorrectly classified images, with the AU 15 region showing generally limited activation. Interestingly, one misclassified example of 'disgust' displayed perfect alignment with expert interpretations. In the case of 'sadness', incorrectly classified images frequently aligned more closely with expert-identified regions. For 'happiness', there was generally a strong correspondence between the activations and expert frameworks; however, some instances showed activations extending beyond the expected areas, a trend consistent in both correctly and incorrectly predicted images. These findings emphasize the complexities and challenges of accurately modeling and interpreting emotional expressions through ML algorithms.

To better understand our AI model's ability to recognize emotional expressions, involving independent experts who have not contributed to the model training or data annotation is essential. These evaluations can confirm if the AI's interpretations align with expert assessments and uncover novel patterns that might escape expert notice. If aligned with expert evaluations, it could indicate that traditional descriptions are missing details that the AI's data-driven methods can identify. Conversely, discrepancies might expose flaws in the AI's training data or fundamental assumptions about emotional expressions. These findings could indicate that integrating expert-defined rules might enhance the AI's accuracy. This critical insight is pivotal for refining AI systems to more accurately emulate human emotional recognition.

Given the complexities revealed in our analysis, we advocate for adopting a Neuro-Symbolic AI approach that integrates data-driven methods (sub-symbolic AI) with domain expert knowledge (symbolic AI). Such hybrid models aim to merge the precision of ML features with the interpretability and dependability of expert-defined rules, significantly enhancing the accuracy and contextual applicability of FER systems. Exploring this further, we identify the following potential technical strategies for implementing Neuro-Symbolic AI systems:

1. Implement a set of universally applicable symbolic rules to establish a broad decision-making framework, supplemented by detailed, data-driven calculations. This strategy harnesses high-level, expert insights to guide the AI's interpretations, enhancing the system's reliability and generalizability.
2. Start with this foundational set of symbolic rules and continuously refine the knowledge base through abductive learning, employing methods outlined in [16]. This approach, as demonstrated in the source, allows the system to dynamically adapt to new or ambiguous emotional expressions and make necessary adjustments - modifying, discarding, or reinforcing rules based on empirical evidence. Such integration not only ensures efficient evolution of the system but also effectively merges theoretical expertise with practical insights to enhance robustness and accuracy in emotion recognition.

By embracing a Neuro-Symbolic approach, we tackle the inconsistencies and limitations observed in purely data-driven models for emotional expression recognition. This methodology not only enhances the system's performance but also sets a foundation for advancing more nuanced and sophisticated AI-driven FER technologies. Such systems could better understand and interpret human emotions in a way that mimics human cognitive processes, offering significant improvements over current models.

This hybrid approach, with its dual reliance on symbolic and sub-symbolic components, promises a richer, more accurate toolkit for developers and researchers aiming to create AI-based FER systems that understand and interact with human emotional states more effectively. Future research could explore the optimization of these strategies, particularly how abductive learning can be fine-tuned to meet specific operational needs or adapt to diverse cultural contexts.

6 Conclusion

In summary, this study examines the use of attribution-based XAI techniques, primarily Grad-CAM combined with Guided Backpropagation, in FER. A key finding is that the features learned by ML models (can) vary within the same emotional state and may not always align with expert interpretations from emotion psychology. This underscores the inherent complexities of using data-driven AI systems, like CNN models, for FER.

To address these challenges, the study proposes integrating data-driven methods (sub-symbolic AI) with domain expert knowledge (symbolic AI) to create Neuro-Symbolic AI systems. By integrating both symbolic and sub-symbolic AI techniques, Neuro-Symbolic AI aims to leverage the advantages of each approach. These hybrid systems can potentially achieve enhanced accuracy, reliability, and explainability by combining learned features from data with expert-defined rules derived from emotion psychology. This integration could facilitate a more nuanced understanding of emotional concepts in FER systems.

References

1. Abhishek, K., Kamath, D.: Attribution-based XAI methods in computer vision: a review. arXiv:2211.14736 (2022)
2. Araf, T.A., Siddika, A., Karimi, S., Alam, M.G.R.: Real-time face emotion recognition and visualization using grad-CAM. In: 2022 Second International Conference on Advances in Electrical, Computing, Communication and Sustainable Technologies (ICAECT), pp. 1–5 (2022). https://doi.org/10.1109/ICAECT54875.2022.9807868
3. Bai, M., Goecke, R., Herath, D.: Micro-expression recognition based on video motion magnification and pre-trained neural network. In: 2021 IEEE International Conference on Image Processing (ICIP), pp. 549–553 (2021). https://doi.org/10.1109/ICIP42928.2021.9506793
4. Barrett, L.F., Adolphs, R., Marsella, S., Martinez, A.M., Pollak, S.D.: Emotional expressions reconsidered: challenges to inferring emotion from human facial movements. Psychol. Sci. Public Interest (2019). https://doi.org/10.1177/1529100619832930
5. Carvalho, D.V., Pereira, E.M., Cardoso, J.S.: Machine learning interpretability: a survey on methods and metrics. Electronics **8**(8), 832 (2019). https://doi.org/10.3390/electronics8080832
6. Chen, G., Zhang, D., Xian, Z., Luo, J., Liang, W., Chen, Y.: Facial expressions classification based on broad learning network. In: 2022 10th International Conference on Information Systems and Computing Technology (ISCTech), pp. 715–720 (2022). https://doi.org/10.1109/ISCTech58360.2022.00118
7. Cheong, J.H., Jolly, E., Xie, T., Byrne, S., Kenney, M., Chang, L.J.: Py-feat: python facial expression analysis toolbox. arXiv:2104.03509 (2023)
8. Deramgozin, M., Jovanovic, S., Rabah, H., Ramzan, N.: A hybrid explainable AI framework applied to global and local facial expression recognition. In: 2021 IEEE International Conference on Imaging Systems and Techniques (IST), pp. 1–5 (2021). https://doi.org/10.1109/IST50367.2021.9651357
9. Ekman, P.: Basic emotions. Handbook of Cognition and Emotion, pp. 301–320. Wiley, New York (1999)
10. Ekman, P., Friesen, W.V., Hager, J.C.: Facial Action Coding System. A Human Face, Salt Lake City, Utah (2002)
11. Fatima, S.A., Kumar, A., Raoof, S.S.: Real time emotion detection of humans using mini-xception algorithm. IOP Conf. Ser. Mater. Sci. Eng. **1042**(1), 012027 (2021). https://doi.org/10.1088/1757-899X/1042/1/012027
12. Gebele, J., Brune, P., Faußer, S.: Face value: on the impact of annotation (in-)consistencies and label ambiguity in facial data on emotion recognition. In: 2022 26th International Conference on Pattern Recognition (ICPR), pp. 2597–2604 (2022). https://doi.org/10.1109/ICPR56361.2022.9956230
13. Gerardo, P.C., Menezes, P.: Classification of FACS-action units with CNN trained from emotion labelled data sets. In: 2019 IEEE International Conference on Systems, Man and Cybernetics (SMC), pp. 3766–3770 (2019). https://doi.org/10.1109/SMC.2019.8914238
14. Goodfellow, I.J., et al.: Challenges in representation learning: a report on three machine learning contests. arXiv:1307.0414 [cs, stat] (2013)
15. Guerdan, L., Raymond, A., Gunes, H.: Toward affective XAI: facial affect analysis for understanding explainable human-AI interactions. In: 2021 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW), pp. 3789–3798 (2021). https://doi.org/10.1109/ICCVW54120.2021.00423

16. Huang, Y.X., Dai, W.Z., Jiang, Y., Zhou, Z.H.: Enabling knowledge refinement upon new concepts in abductive learning. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 37, pp. 7928–7935 (2023)
17. Kishan Kondaveeti, H., Vishal Goud, M.: Emotion detection using deep facial features. In: 2020 IEEE International Conference on Advent Trends in Multidisciplinary Research and Innovation (ICATMRI), pp. 1–8 (2020). https://doi.org/10.1109/ICATMRI51801.2020.9398439
18. Krizhevsky, A., Sutskever, I., Hinton, G.E.: ImageNet classification with deep convolutional neural networks. In: Advances in Neural Information Processing Systems, vol. 25, pp. 1097–1105 (2012)
19. Li, S., Deng, W.: Deep facial expression recognition: a survey. IEEE Trans. Affect. Comput. **13**(3), 1195–1215 (2022). https://doi.org/10.1109/TAFFC.2020.2981446
20. Li, S., Deng, W., Du, J.: Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2584–2593. IEEE, Honolulu, HI (2017). https://doi.org/10.1109/CVPR.2017.277
21. Lucey, P., Cohn, J.F., Kanade, T., Saragih, J., Ambadar, Z., Matthews, I.: The extended cohn-kanade dataset (CK+): a complete dataset for action unit and emotion-specified expression. In: 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Workshops, pp. 94–101 (2010). https://doi.org/10.1109/CVPRW.2010.5543262
22. Lundberg, S., Lee, S.I.: A unified approach to interpreting model predictions (2017). https://doi.org/10.48550/arXiv.1705.07874
23. Mahendran, A., Vedaldi, A.: Visualizing deep convolutional neural networks using natural pre-images. Int. J. Comput. Vision **120**(3), 233–255 (2016). https://doi.org/10.1007/s11263-016-0911-8
24. Malek–Podjaski, M., Deligianni, F.: Towards explainable, privacy-preserved human-motion affect recognition. In: 2021 IEEE Symposium Series on Computational Intelligence (SSCI), pp. 01–09 (2021). https://doi.org/10.1109/SSCI50451.2021.9660129
25. Mollahosseini, A., Hasani, B., Mahoor, M.H.: AffectNet: a database for facial expression, valence, and arousal computing in the wild. IEEE Trans. Affect. Comput. **10**(1), 18–31 (2017)
26. Molnar, C.: Interpretable Machine Learning (Second Edition) - A Guide for Making Black Box Models Explainable. Leanpub (2018)
27. Moreno-Armendáriz, M.A., Espinosa-Juarez, A., Godinez-Montero, E.: Using diverse ConvNets to classify face action units in dataset on emotions among Mexicans (DEM). IEEE Access **12**, 15268–15279 (2024)
28. Mouakher, A., Chatry, S., Yacoubi, S.E.: A multi-criteria evaluation framework for facial expression recognition models. In: 2023 20th ACS/IEEE International Conference on Computer Systems and Applications (AICCSA), pp. 1–8 (2023). https://doi.org/10.1109/AICCSA59173.2023.10479285
29. Ribeiro, M.T., Singh, S., Guestrin, C.: "Why should I trust you?": explaining the predictions of any classifier. arXiv:1602.04938 (2016)
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. Int. J. Comput. Vision **128**(2), 336–359 (2020). https://doi.org/10.1007/s11263-019-01228-7
31. Shahabinejad, M., Wang, Y., Yu, Y., Tang, J., Li, J.: Toward personalized emotion recognition: a face recognition based attention method for facial emotion recognition. In: 2021 16th IEEE International Conference on Automatic Face and Gesture

Recognition (FG 2021), pp. 1–5 (2021). https://doi.org/10.1109/FG52635.2021.9666982
32. Shingjergji, K., Iren, D., Böttger, F., Urlings, C., Klemke, R.: Interpretable explainability in facial emotion recognition and gamification for data collection. In: 2022 10th International Conference on Affective Computing and Intelligent Interaction (ACII), pp. 1–8 (2022). https://doi.org/10.1109/ACII55700.2022.9953864
33. Simonyan, K., Vedaldi, A., Zisserman, A.: Deep inside convolutional networks: visualising image classification models and saliency maps. arXiv:1312.6034 (2014)
34. Smilkov, D., Thorat, N., Kim, B., Viégas, F., Wattenberg, M.: SmoothGrad: removing noise by adding noise. arXiv:1706.03825 (2017)
35. Wadhawan, R., Gandhi, T.K.: Landmark-aware and part-based ensemble transfer learning network for static facial expression recognition from images. IEEE Trans. Artif. Intell. **4**(2), 349–361 (2023). https://doi.org/10.1109/TAI.2022.3172272
36. Yang, J., Zhang, F., Chen, B., Khan, S.U.: Facial expression recognition based on facial action unit. In: 2019 Tenth International Green and Sustainable Computing Conference (IGSC), pp. 1–6 (2019). https://doi.org/10.1109/IGSC48788.2019.8957163
37. Zeiler, M.D., Fergus, R.: Visualizing and understanding convolutional networks. arXiv:1311.2901 (2013)
38. Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., Torralba, A.: Learning deep features for discriminative localization. arXiv:1512.04150 (2015)

Securing Faces: A GAN-Powered Defense Against Spoofing with MSRCR and CBAM

Aashania Antil and Chhavi Dhiman(✉)

Delhi Technological University, Delhi, India
chhavi.dhiman@dtu.ac.in

Abstract. Ensuring the security of face authentication systems is crucial, and Face Anti-Spoofing System (FAS) play a key role in defending against spoofing threats. Depth-supervised learning has proven effective in FAS, utilizing depth maps as auxiliary features due to their computational simplicity. However, existing methods often struggle to generalize effectively in intricate environments and counter unknown attacks. To address this challenge, our work introduces a novel GAN-based architecture for FAS. To enhance generalization, we introduce Multi-Scale Retinex with Color Restoration (MSRCR) images alongside RGB, and apply the Convolutional Block Attention Module (CBAM) mechanism within the generator framework to highlight salient features. The classifier is trained using a latent variable encompassing depth information, improving generalization across diverse environmental conditions, including variations in illumination and background. Experimental results demonstrate the effectiveness of our approach, outperforming other methods on multiple datasets including CASIA-FASD, MSU-MFSD, OULU-NPU and Replay-Attack for both intra-dataset and cross-dataset testing between Replay-Attack and CASIA-FASD datasets.

Keywords: Face anti-spoofing · presentation attack detection · generative adversarial network · multi-scale retinex with color restoration

1 Introduction

The rise of face recognition (FR) technology in interactive AI systems has revolutionized human-computer interaction, offering convenience and unparalleled accuracy. However, the widespread adoption of FR systems has introduced a vulnerability to presentation attacks (PAs) like printed images [1], video replays [2], and 3D masks [3]. These deceptive methods pose a serious threat to the reliability of face authentication systems. Recognizing the pressing need for enhanced security, the advancement of face anti-spoofing (FAS) approaches has become critical, focusing on detecting and mitigating PAs.

Over the last two decades, presentation attack detection (PAD) methods have evolved from traditional to deep learning-based approaches. Traditional methods primarily rely on hand-crafted operators [4] for feature extraction in PAD. However, these methods are susceptible to environmental variations, leading to suboptimal generalization performance and impracticality. Alternatively, certain deep learning-based methods focus on

A. Antonacopoulos et al. (Eds.): ICPR 2024, LNCS 15313, pp. 430–449, 2025.
https://doi.org/10.1007/978-3-031-78201-5_28

face liveness detection at the frame and video levels. Many of these approaches frame FAS as a binary classification problem [5, 6], supervised by a binary cross-entropy loss. However, CNNs using binary loss encounter difficulties in discerning authentic spoofing patterns. The introduction of pixel-wise supervision, incorporating labels such as binary masks [7], pseudo depth [8] etc., has improved context-aware signals for local live/spoofing cues. Despite these improvements, current implementations remain susceptible to disturbances. Existing FAS methods often overfit specific scenarios, making them vulnerable to domain shifts and unforeseen attacks. Strategies like multi-domain disentangled [9] learning aim to enhance generalization. However, detailed spoofing pattern learning remains challenging, particularly with coarse supervision signals.

Lately, Generative Adversarial Networks (GANs) [9, 10] have significantly advanced generative modeling, excelling in representation learning and domain-to-domain data transformation. Their efficacy arises from the incorporation of a discriminator, framing the learning problem as a two-player minimax game. In FAS, Wang et al. [11] introduced an innovative approach— the domain transfer network (DTN) employing a GAN. This approach effectively shifts the domain from RGB to depth, allowing for the seamless fusion of latent feature embeddings from both sources. Consequently, it achieved significantly enhanced generalizability compared to existing methods. Inspired by the successes of GANs, this paper explores the effectiveness of depth data for generating high-quality information in FAS. To realize this, we introduce a novel framework that leverages the capabilities of GANs to convert input RGB images into depth maps for live/spoof classification. Our proposed framework sets itself apart by integrating Multi-Scale Retinex with color restoration filtering (MSRCR) [12] alongside RGB, enhancing the input data before utilizing it as the main input for the generator. Both RGB and MSRCR contribute discriminative information that is particularly effective for spoofing detection. To fully leverage the strengths of both, we employ the Convolutional Block Attention Module (CBAM) [13] to accentuate essential features, which prove indispensable for generating precise depth maps within the generative network. Following this framework, our generative network adeptly captures distinctive features and channels them to the auxiliary classifier for binary classification.

The significant contributions of this paper can be outlined as:

- We propose a novel GAN-based framework for face anti-spoofing that integrates MSRCR along with RGB to enhance input data quality and provide discriminative information to the framework.
- We incorporate the CBAM module to enhance the attention, facilitating precise depth map generation using generative network.
- We conducted comprehensive experiments on the CASIA-FASD [1], MSU-MFSD [14], Replay-Attack [2], and OULU-NPU [15] datasets for intra- and cross-dataset testing. All experiments results show that proposed framework performs comparably with the other state-of-the-arts.

The paper is divided into these sections: Sect. 2 gives a concise review of relevant research, Sect. 3 elaborates on the proposed framework, Sect. 4 reports on extensive experiments that demonstrate its effectiveness, and Sect. 5 concludes the paper.

2 Related Work

Face Anti-spoofing. FAS methodologies are generally categorized into three main types: Traditional, Deep learning, and Generalized feature learning-based methods. Traditional methods rely on extracting handcrafted features [16, 17] from facial images, aiming to capture spoof patterns [4] through color, texture, motion, and liveliness cues. The advent of deep learning in computer vision has led to diverse approaches [9], treating FAS as a binary classification task using binary supervision or adopting pixel-wise supervisions like pseudo depth maps [8], reflection maps [18] and binary maps [7] for fine-grained learning. Researchers employ a variety of backbones, including CNNs [19], pre-trained VGG16 [20], lightweight networks, and vision transformers [5, 21, 22]. However, the scarcity of diverse and large-scale datasets for spoof attacks poses challenges, leading to overfitting and vulnerability to unseen attacks and domain shift [23] in existing FAS methods. To address this, Generalized feature learning-based approaches have emerged, focusing on generalization to unseen scenarios. Techniques such as domain adaptation [24, 25] and domain generation [26, 27] have been developed, emphasizing the potential of methodologies based on generative models. Yet, challenges persist, including the need for domain labels and the limitations associated with binary classification. The ongoing pursuit of robust and adaptable FAS methods continues to navigate these complexities.

Generative Adversarial Network. GANs showcase remarkable versatility in image generation tasks. In particular, the dual encoder-decoder GAN [28] specializes in creating facial images with seamless pose transitions, while the self-growing and pruning GAN [29] enhances network stability and image quality. GANs excel in tasks involving image-to-image transfer [30], highlighting their adaptability across diverse domains. Within the realm of FAS, GAN-based methods [10, 11, 31] are making significant strides in distinguishing between genuine and spoof faces. For example, STDN [32] integrates a disentanglement generator, a reconstruction & synthesis module, and a multi-scale approach, though it grapples with challenges related to unpaired samples. DC-MS [33] employs feature swapping for feature-level decoupling, albeit with computation-intensive processes. AIM-FAS [34] addresses FAS as a zero/few-shot learning problem, while CMA [35] leverages cross-modal transfer for style transformation from RGB to NIR modality. Additionally, a semi-supervised framework [36] reduces reliance on annotated data using pseudo labels. These advancements collectively underscore the substantial potential of GAN-based techniques in significantly enhancing FAS methodologies.

3 Proposed Methodology

The schematic diagram in Fig. 1 illustrates the novel GAN-based FAS framework, which integrates three essential components: a generator ($\boldsymbol{G}$), a discriminator ($\boldsymbol{D}$), and a classifier ($\boldsymbol{C}$). The generator ($\boldsymbol{G}$) maps RGB images to depth maps, the discriminator ($\boldsymbol{D}$) evaluates the accuracy of these maps, and the classifier ($\boldsymbol{C}$) differentiates between genuine and spoofed faces.

Using characteristics extracted solely from RGB images is not consistently effective in distinguishing between live and spoofed faces, the generator employs a dual branch

architecture. One branch process raw RGB images directly, while the other converts RGB images into Multi-Scale Retinex with Color Restoration (MSRCR) images followed by Convolutional Block Attention Module (CBAM). CBAM enhances essential features by considering their significance across different channels and locations. RGB images capture detailed facial textures but are sensitive to illumination variations. Conversely, MSRCR images maintain invariance to illumination changes, albeit with a loss of some minor facial details. To leverage the strengths of both branches, the framework performs element-wise multiplication on the features obtained from each branch. This combined feature is then fed into a U-Net architecture to generate depth maps. Following the GAN paradigm, the discriminator classifies real and fake depth images, helping the generator improve its outputs over time. The classifier, illustrated in Fig. 1, produces a binary output indicating whether the input face image is live or spoofed. It processes feature maps generated by generator, treating them as depth domain features.

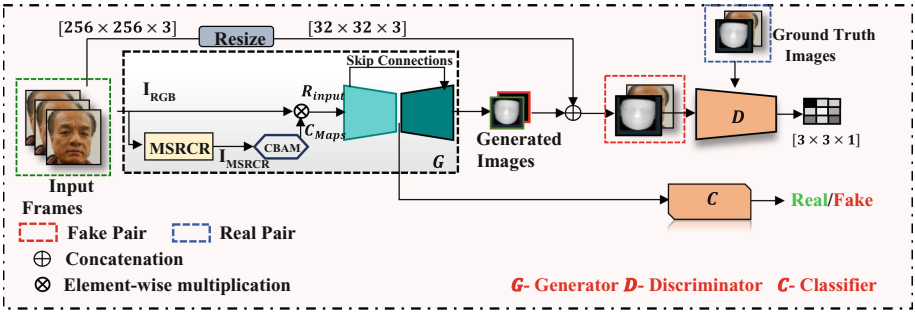


Fig. 1. Illustration of the overall proposed GAN-based framework.

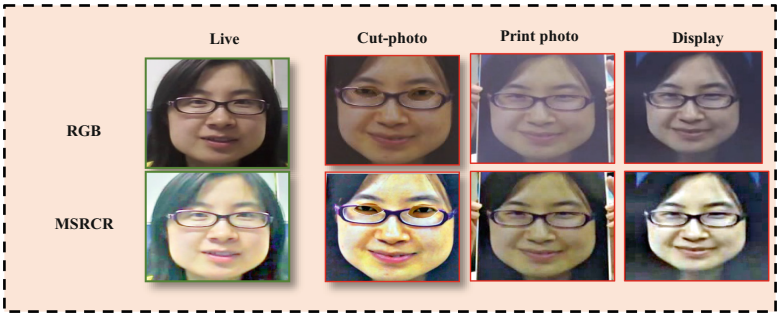


Fig. 2. CASIA-FASD images with pre-processing: green boundaries represent live samples and red boundaries indicate different types of attacks. Top and bottom: RGB and MSRCR. (Color figure online)

This section begins by introducing the MSRCR and CBAM blocks, then provides a detailed explanation of the proposed network framework.

3.1 Multi-scale Retinex with Color Restoration (MSRCR)

Various studies have explored algorithms that simulate the human visual system, with a primary focus on luminance. Land's Retinex theory [12] introduced a lightness model that proved successful for image enhancement [37]. The Single Scale Retinex (SSR) model further refined this approach [38], utilizing a Gaussian filter to normalize image illumination. In a subsequent enhancement to SSR, [39] incorporated a guided filter, yielding promising results in image enhancement. Building upon the SSR foundation, the Multi-Scale Retinex (MSR) model [12] amalgamates outputs to create a comprehensive image enhancement methodology. For contrast enhancement, [40] an adaptive-weight MSR was introduced, normalizing output pixel values within the range of [0, 255]. Addressing color deviation in MSR, Jobson et al. [12] proposed a solution by multiplying the MSR result with a Color Restoration (CR) function. The MSRCR function includes gain and offset for each channel, effectively enhancing image contrast. Applied independently to each color channel, MSRCR is particularly adept at addressing images influenced by colored illumination. Our evaluation on diverse datasets underscores MSRCR's adaptability and notable performance improvements, enhancing color information under varying lighting conditions. This study provides valuable insights into MSRCR's strength in real-world scenarios with colored illumination. In Fig. 2, images from the CASIA-FASD [1] dataset are depicted, presenting live and attack samples in RGB and MSRCR representations.

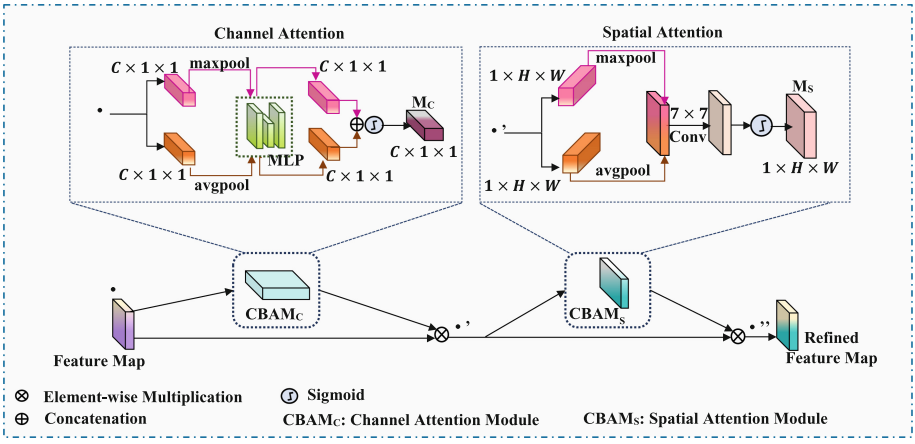


Fig. 3. Schematic diagram of CBAM Module [13]

3.2 Convolutional Block Attention Module (CBAM)

The Sequential Channel and Spatial Attention Module plays a pivotal role in highlighting salient features embedded within the Generator framework. This module enhances feature representation by creating a refined feature map, $\mathbb{F}''$, through the processing of MSRCR images (I_{MSRCR}) obtained from RGB inputs, I_{RGB}. When I_{MSRCR} images enter the channel attention block, they are denoted as $\mathbb{F}\epsilon\mathbb{R}^{C\times H\times W}$:. This process compresses the spatial dimension, yielding a 1D channel attention map $M_C\epsilon\mathbb{R}^{C\times 1\times 1}$. Subsequently, the modified tensor $\mathbb{F}'$, as given in Eq. (1), passes to the spatial attention block, resulting in a two-dimensional attention map $M_s\epsilon\mathbb{R}^{1\times H\times W}$. The final output, $\mathbb{F}''$, is expressed by Eq. (2):

$$\mathbb{F}' = CBAM_C(\mathbb{F}) \otimes (\mathbb{F}) \tag{1}$$

$$\mathbb{F}'' = CBAM_S(\mathbb{F}') \otimes (\mathbb{F}') \tag{2}$$

here $\otimes$ signifies the element-wise multiplication. CBAM's architecture, depicted in Fig. 3, involves two pivotal processes:

- ***Computation of Channel Attention***: This process enhances finer details and reduces information loss by simultaneously applying average and max pooling to aggregate the squeezed spatial dimensions. The resulting descriptors, $\mathbb{F}^c_{avg}$and $\mathbb{F}^c_{max}$, are fed into a shared MLP (Multilayer Perceptron) with a single hidden layer. The shared layer uses a reduction ratio of 8 to minimize computation and parameter overhead. The MLP outputs are combined element-wise, followed by the application of sigmoid function resulting in the channel attention map $M_c(\mathbb{F})\epsilon\mathbb{R}^{C\times 1\times 1}$, as specified in Eq. (3):

$$\begin{aligned} M_C(\mathbb{F}) &= \sigma\big(MLP(maxpool(\mathbb{F})) + MLP(avgpool(\mathbb{F}))\big) \\ &= \sigma(W_1(W_0(\mathbb{F}^c_{max})) + W_1(W_0(\mathbb{F}^c_{avg}))) \end{aligned} \tag{3}$$

here, the symbol " + " represents the element-wise addition.

- ***Computation of Spatial Attention:*** This process emphasizes important details within the feature map by utilizing max and average pooling across the channel dimensions to produce 2D feature maps, $\mathbb{F}^s_{max}\epsilon\mathbb{R}^{1\times H\times W}$and $\mathbb{F}^s_{avg}\epsilon\mathbb{R}^{1\times H\times W}$. These feature representations are then aggregated to form a robust map, which is convolved with 7 × 7 filter size to produce the spatial attention map $M_s(\mathbb{F})\epsilon\mathbb{R}^{1\times H\times W}$described as follows:

$$\begin{aligned} M_s(\mathbb{F}) &= \sigma(f^{(7x7)}\{maxpool(\mathbb{F}); avgpool(\mathbb{F})\}) \\ &= \sigma(f^{(7x7)}\{\mathbb{F}^s_{max}; \mathbb{F}^s_{avg}\}) \end{aligned} \tag{4}$$

here, σ represents sigmoid function, and f_{7x7} denotes a convolutional operation with a 7 × 7 filter size.

Equations (3) and (4) collectively describe a feature map refinement process, operating along both the channel and spatial axes, respectively. The resultant CBAM feature map, C_{Maps} showcases augmented representation in both dimensions, thereby enhancing the network's feature extraction capabilities.

3.3 Network Architecture

Generator. As shown in Fig. 1, the generator, $\boldsymbol{G}$ uses an encoder-decoder architecture with EfficientNetB4 [41] as the backbone, achieving a balance between model capacity and size. The encoder encodes RGB images for the decoder to exclusively reconstruct depth maps. Our framework prioritizes multi-scale features and incorporates skip connections to form a U-Net architecture. These connections seamlessly merge features from both the encoder and up-sampling segments to ensure a smooth flow of information, improving gradient flow and effectively tackling the vanishing gradient issue, particularly in lower layers.

Despite the powerful non-linear feature learning of deep learning, anti-spoofing performance degrades with varying input conditions. To address this, we introduce MSRCR images (I_{MSRCR}) obtained by converting RGB images (I_{RGB}). MSRCR operates on the intensity and chromatic channels separately, achieving both illumination invariance and color fidelity. Unlike the commonly used RGB color space, prone to illumination sensitivity, MSRCR separates illumination information from color details. This creates a representation resilient to lighting changes enhancing discriminative information for spoofing detection and forming a complementary relationship with the detailed yet illumination-sensitive RGB representation. To maximize the strengths of both, we employ a CBAM mechanism specifically on I_{MSRCR}, highlighting essential features based on the context and considering their significance across different channels and locations. The resulting attention map, C_{Maps}, provides weight information for each pixel in an image. To optimize this information, we perform element-wise multiplication between the obtained C_{Maps} and I_{RGB}, generating the final refined input, R_{input}, expressed as:

$$R_{input} = I_{RGB} \otimes C_{Maps} \tag{5}$$

This R_{input} is then fed as the final input to the U-Net-based architecture, ensuring the meaningful extraction of features passed on to the decoder blocks. Following the last decoder block, depth maps are generated using a convolution layer and a Tanh activation.

Discriminator. In the proposed framework, the discriminator $\boldsymbol{D}$ follows a PatchGAN-inspired architecture [42], consisting of multiple convolutional layers, each followed by batch normalization and LeakyReLU activation. $\boldsymbol{D}$ processes two inputs of dimensions $[32 \times 32]$, comprising a real pair with ground truth depth maps (D) associated with I_{RGB}, and a fake pair composed of generated depth maps $\{\mathrm{D} = \boldsymbol{G}(\mathrm{I})\}$paired with I_{RGB}. This adversarial training allows $\boldsymbol{G}$ to improve gradually, guided by $\boldsymbol{D}$'s gradients, enhancing its performance in generating highly realistic depth maps from I_{RGB}. The framework's objective is defined as:

$$\mathbb{L}_{GAN}(\boldsymbol{G}, \boldsymbol{D}) = \mathbb{E}_{I,D}[log\boldsymbol{D}(I, D)] + \mathbb{E}_{I}[log(1 - \boldsymbol{D}(I, \mathrm{D})] \tag{6}$$

where, D represents ground truth depth maps, and $\mathrm{D} = \boldsymbol{G}(I)$. To ensure network stability, conventional methods incorporate supplementary image reconstruction losses (L1 or L2 distance). With the L1 reconstruction loss:

$$\mathbb{L}_{L1}(\boldsymbol{G}) = \mathbb{E}_{I,D}[\big|\big|D - \mathrm{D}\big|\big|_1] \tag{7}$$

The proposed framework is optimized using the combined objective:

$$\mathbb{L}_{our}(\boldsymbol{G}, \boldsymbol{D}) = arg_{\boldsymbol{G}}^{min}{}_{\boldsymbol{D}}^{max}\mathbb{L}_{GAN}(\boldsymbol{G}, \boldsymbol{D}) + \lambda\mathbb{L}_{L1}(\boldsymbol{G}) \tag{8}$$

where λ represents a balancing parameter, ensuring effective training for detailed and accurate depth image generation.

Classifier. In the FAS task, the classifier is the final component responsible for distinguishing between live and spoofed faces. It takes input from the latent variable within the generator's encoder, presumed to contain depth-representing features post-training in the GAN network. The GAN training process ensures the effective integration of RGB and depth features by the encoder, enhancing the classifier's overall generalization. The classifier is optimized using cross-entropy, guiding the model to identify subtle features in the latent variable for accurate image classification. This seamless integration of learned representations from the encoder significantly strengthens the FAS system's classifier. The classifier's loss function is expressed as:

$$\mathbb{L}_C = -(ylog(p) + (1-y)log(1-p)) \tag{9}$$

here p depicts the predicted probability and y represents the ground truth label. This formulation captures the essence of the classifier's task in evaluating the probability of a given input image being either genuine or spoofed.

4 Experiments

In this section, we provide detailed insights into the datasets used in our study and the metrics employed to evaluate our models. We describe our experimental setup, highlight our achieved results, compare them with the latest benchmarks specific to each dataset, and discuss our analysis of performance variations through an ablation study.

4.1 Datasets

To assess the efficacy and versatility of our framework, we conduct rigorous evaluations on four benchmark datasets: MSU-MFSD [14], CASIA-FASD [1], Replay-Attack [2], and OULU-NPU [15]. MSU-MFSD (M) [14] contains 280 video clips showcasing various photo and video-based attacks on 35 clients, categorized into three distinct spoof attack types. CASIA-FASD (C) [1] comprises 600 videos with real and spoof faces, featuring diverse attacks and varying image qualities captured from three types of cameras. We split the dataset into training and testing sets, with 20 and 30 subjects, respectively, for comprehensive evaluation. The REPLAY-ATTACK (RA) [2] dataset includes 1200 videos with genuine and spoofed faces, considering different illumination and support conditions. We divide it into training, development, and testing sets for comprehensive evaluation. Lastly, OULU-NPU (O) [15] consists of 5940 videos documenting genuine access attempts and attacks across different contexts, with a split into training, development, and testing subsets. This thorough evaluation on diverse datasets ensures a comprehensive assessment of our proposed approach.

4.2 Performance Metrics

For a comprehensive comparison with prior research, we employ specific evaluation metrics corresponding to each benchmark dataset. In the C dataset, model outcomes are evaluated using the Equal Error Rate (EER) on the test set. For the RA benchmark, we adopt the Half Total Error Rate (HTER). In the O dataset, we employ the ISO/IEC 30107–3 metrics [43], which include Attack Presentation Classification Error Rate (APCER), Bona Fide Presentation Classification Error Rate (BPCER), and Average Classification Error Rate (ACER). These metrics correspond to APCER, BPCER, and their average, respectively. To ensure generalizability across datasets, our primary evaluation criterion involves HTER for cross-dataset testing between C and RA.

4.3 Implementation Details

Data Preprocessing. Our data preprocessing follows established research practices and involves several steps to ensure consistency and accuracy across the benchmark datasets. For video benchmarks, we employ frame sampling, face alignment, and systematic extraction of frames at a 10th frame interval. Initially, face detection utilizes the Viola-Jones algorithm [44], later transitioning to MTCNN [45] due to observed limitations in specific datasets. Ground-truth depth maps are generated using PRNet [46], resulting in [32 $\times$ 32] depth maps for genuine live faces and zeros for spoofed samples. Our methodology includes additional steps such as random horizontal flipping and data augmentation to enhance dataset diversity.

Training Setup. The proposed framework is implemented using Keras and experiments are conducted in the Google Colab Pro environment with the support of an Nvidia T4 GPU and 16 GB of RAM. The generator backbone, EfficientNetB4 [41], is initialized using a pre-trained model from ImageNet. Newly introduced modules follow the "He-Uniform" initialization approach. Network optimization employs the Adam optimizer, initialized with a learning rate of $1e-3$, and a batch size of 16. The loss function configuration includes λ_{GAN} and λ_{L1} set to 1 and 100, respectively [11]. During each training epoch, images are randomly shuffled and flipped, enhancing the diversity of the training dataset.

4.4 Comparison with Other State-of-the-Arts

Intra-dataset Testing. In the intra-dataset testing, the training and testing sets are derived from the same datasets to evaluate the performance of our framework in face PA detection. We evaluated our method against leading approaches using the CASIA-FASD (C), Replay-Attack (RA), and OULU-NPU (O) datasets, adhering to established protocols and benchmarking against state-of-the-arts [16, 24, 47–51].

Table 1 presents the spoofing detection results in terms of EER values for the C dataset. Our framework surpasses all statistical methods and achieves the second-highest performance with an EER of 1.21% following Zhang et al. [31] with 1.17%. Zhang et al. [31] utilized Wasserstein loss and incorporated shortcut connections within the generator. For the RA dataset, Table 2 provides both EER and HTER values. Our framework delivers the highest performance with respect to EER (0.05%) and secures the second-best HTER at 0.03%, with only a marginal difference of 0.01% compared to the DTN [11] (0.02%), which is computationally expensive. This demonstrates the framework's robust capability in distinguishing between genuine and spoof faces, attributable to the integration of RGB and MSRCR features within the generator.

The comparative results for the OULU-NPU dataset are given in Table 3. In Protocol 1, although SGTD [55] and Zhang et al. [31] exhibit superior BPCER, our model achieves a low ACER of 0.6%. Under Protocol 2, our approach performs comparably to CDCN [56], ranking second in both APCER and ACER. In Protocol 3, our method leads in APCER and ACER, securing the third position in BPCER, following STASN [57] and CDCN [56]. For Protocol 4, our approach achieves the highest performance in APCER and the second-highest in BPCER and ACER, like the DTN [11] approach. These results validate the generalization capability and effectiveness of our framework, particularly as Protocol IV evaluates the model across all challenging aspects of the database. The intra-dataset evaluation shows that our framework delivers strong effectively, yielding competitive results across the benchmark datasets.

Table 1. Intra-Dataset evaluation on CASIA-FASD(C) Dataset (%).

Methods	EER
Patch and Depth [48]	2.67
LBP [16]	18.2
Color Texture [47]	6.20
Attention [49]	3.14
ML-DAN [24]	3.7
FARCNN [50]	2.35
DTN [11]	1.34
MIQF-SVM [52]	12.7
DOG-ADTCP [53]	–
Zhang et al. [31]	**1.17**
DSCNN [54]	2.9
Ours	**1.21**

Table 2. Intra-Dataset evaluation on Replay-Attack (RA) Dataset (%).

Methods	EER	HTER
Patch and Depth [48]	0.79	0.72
LBP [16]	13.9	13.8
Color Texture [47]	0.4	2.9
Attention [49]	0.13	0.25
ML-DAN [24]	0.3	0.6
FARCNN [50]	0.06	0.18
DTN [11]	0.06	**0.02**
MIQF-SVM [52]	–	5.38
DOG-ADTCP [53]	0.81	3.24
Zhang et al. [31]	0.09	0.22
DSCNN [54]	4.7	0.39
Ours	**0.05**	**0.03**

Cross-dataset Testing. The adaptability of a FAS framework across diverse environments is important for practical applications. To assess the generalization capability of our proposed framework, we initially conducted cross-dataset testing using two labeled datasets, referred to as C and RA. The outcomes, presented in terms of HTER in Table 4, were obtained through two distinct experimental protocols. The first protocol involved training on the C dataset and testing on the RA dataset, while the second protocol reversed this arrangement. Table 4 demonstrates the excellence of our framework in both scenarios.

To delve deeper into the framework's generalization ability, we performed cross-dataset testing across four different datasets, producing four unique test cases, as shown in Table 5. In each test case, a single dataset was chosen as the testing set, with the other three used for training. The four test cases were as follows: Test Case 1 - O&C&RA to M, Test Case 2 - O&M&RA to C, Test Case 3 - O&C&M to RA, and Test Case 4 - RA&C&M to O. The results from these test cases indicate that our framework demonstrates superior performance in Test Case 1, outperforming other methods. In Test Case 2, it ranks second, just behind CDCN-PS [58]. For Test Case 3, our framework ranks fifth, showing inferior performance compared to CDCN [56], which is based on contrastive learning. In Test Case 4, our framework ranks third, with a marginal difference from CDCN [56] and CDCN-PS [58]. These findings highlight the capability of our GAN-based approach in advancing the field of FAS.

Table 3. Intra-Dataset evaluation on OULU-NPU (O) (%).

Protocols	Methods	APCER	BPCER	ACER
1	CPqD [59]	2.9	10.8	6.9
	Auxiliary [60]	1.6	1.6	1.6
	FaceDs [61]	1.2	1.7	1.5
	STASN [57]	1.2	2.5	1.9
	CDCN [56]	0.4	1.7	1.1
	SGTD [55]	2	**0**	1
	DTN [11]	0.78	1.06	0.92
	MIQF-SVM [52]	6.9	1.5	4.2
	Zhang et al. [31]	0.63	0.80	0.72
	DSCNN [54]	0.37	2.9	1.6
	Ours	**0.3**	**0.9**	**0.6**
2	CPqD [59]	14.7	3.6	9.2
	Auxiliary [60]	2.7	2.7	2.7
	FaceDs [61]	4.2	4.4	4.3
	STASN [57]	4.2	0.3	2.2
	CDCN [56]	1.5	1.4	1.5
	SGTD [55]	2.5	1.3	1.9
	DTN [11]	3.84	2.11	2.88
	MIQF-SVM [52]	7.8	1.4	4.6
	Zhang et al. [31]	2.53	1.36	1.95
	DSCNN [54]	3.1	7.2	5.2
	Ours	**2.5**	**1.1**	**1.8**
3	CPqD [59]	6.8 ± 5.6	8.1 ± 6.4	7.4 ± 3.3
	Auxiliary [60]	2.7 ± 1.3	3.1 ± 1.7	2.9 ± 1.5
	FaceDs [61]	4.0 ± 1.8	3.8 ± 1.2	3.6 ± 1.6
	STASN [57]	4.7 ± 3.9	0.9 ± 1.2	2.8 ± 1.6
	CDCN [56]	2.4 ± 1.3	2.2 ± 2.0	2.3 ± 1.4
	SGTD [55]	3.2 ± 2.0	2.2 ± 1.4	2.7 ± 0.6
	DTN [11]	1.9 ± 1.6	3.8 ± 6.4	2.8 ± 2.7
	MIQF-SVM [52]	3.6 ± 0.9	4.3 ± 1.8	4.0 ± 1.4
	Zhang et al. [31]	1.7 ± 1.4	2.7 ± 4.3	2.2 ± 3.0
	DSCNN [54]	5.6 ± 1.7	4 ± 3.3	4.8 ± 2.5

(*continued*)

Table 3. (*continued*)

Protocols	Methods	APCER	BPCER	ACER
	Ours	**1.6 ± 1.1**	**2.5 ± 1.0**	**2.05 ± 1.1**
4	CPqD [59]	32.5 ± 37.5	11.7 ± 12.1	22.1 ± 20.8
	Auxiliary [60]	9.3 ± 5.6	10.4 ± 6.0	9.5 ± 6.0
	FaceDs [61]	1.2 ± 6.3	6.1 ± 5.1	5.6 ± 5.7
	STASN [57]	6.7 ± 10.6	8.3 ± 8.4	7.5 ± 4.7
	CDCN [56]	4.6 ± 4.6	9.2 ± 8.0	6.9 ± 2.9
	SGTD [55]	6.7 ± 7.5	3.3 ± 4.1	5.0 ± 2.2
	DTN [11]	4.0 ± 4.1	3.0 ± 4.9	3.5 ± 2.4
	MIQF-SVM [52]	6.2 ± 4.3	4.9 ± 3.7	5.6 ± 4.0
	Zhang et al. [31]	2.1 ± 4.5	5.7 ± 4.9	3.9 ± 3.2
	DSCNN [54]	9.6 ± 6.0	7.8 ± 5.6	9.3 ± 6.3
	Ours	**3.8 ± 2.5**	**3.2 ± 4.6**	**3.5 ± 3.5**

Table 4. Comparative Analysis of Cross-Dataset Testing CASIA-FASD vs. Replay-Attack in terms of HTER (%)

Methods	Train- C/Test- RA	Train- RA/Test-C
LBP [16]	55.9	47.9
LBP-TOP [62]	49.7	60.6
Color Texture [47]	47.0	39.6
Deep-Learning [63]	48.2	45.4
Auxiliary [60]	27.9	28.4
STASN [57]	31.5	30.9
FARCNN [50]	26.0	29.4
Attention [49]	30.0	33.4
DTN [11]	16.64	22.98
Zhang et al. [31]	25.73	21.57
Ours	**15.8**	**21.17**

Table 5. Comparative Analysis of Cross-Dataset Testing Across Four Datasets (%)

Methods	Test Case 1		Test Case 2		Test Case 3		Test Case 4	
	HTER	AUC	HTER	AUC	HTER	AUC	HTER	AUC
LBP-TOP [62]	36.9	70.80	42.6	61.05	49.45	49.54	53.15	44.09
MMD-AAE [64]	27.08	83.19	44.59	58.29	31.58	75.18	40.98	63.08
Color Texture [4]	28.09	78.47	30.58	76.89	40.4	62.78	63.59	32.71
Binary CNN [19]	29.25	82.87	34.88	71.94	34.47	65.88	29.61	77.54
MADDG [65]	17.69	88.06	24.5	84.51	22.19	84.99	27.98	80.02
Auxiliary [60]	22.72	85.88	33.52	73.15	29.14	71.69	30.17	77.61
CDCN [56]	22.90	85.45	22.46	86.64	19.98	84.75	16.92	90.46
CDCN-PS [58]	20.42	87.43	18.25	86.76	19.55	86.38	15.76	92.43
DTN [51]	19.40	86.87	22.03	87.71	21.43	88.81	18.26	89.40
Ours	16.3	90.7	21.17	83.55	23.21	85.7	17.65	89.69

Test Case 1: C&RA&O to M; **Test Case 2**: M&RA&O to C; **Test Case 3:** C&M&O to RA; **Test Case 4**: C&RA&M to O.

Table 6. Ablation Study Results: OULU-NPU Protocol 2 (%)

Backbone	Input	APCER	BPCER	ACER
Simple Encoder-Decoder (w/o skip connections)	RGB	11.5	8.9	10.2
U-Net	RGB	9.6	7.7	8.65
	RGB + MSRCR	5.4	4.8	5.1
With CBAM	RGB + MSRCR	2.5	1.1	1.8

4.5 Ablation Study

Our ablation studies use the OULU-NPU protocol 2 as the exclusive testing benchmark, focusing on component selection within our framework. Initial trials with a basic encoder-decoder network yielded in a high ACER of 10.2% due to challenges in accurate depth map generation from its bottleneck design. Subsequently, adopting the U-Net architecture with skip connections significantly improved performance reduced the ACER. As shown in Table 6, standard RGB input led to 8.65% ACER, while introducing MSRCR alongside RGB input lowered the ACER to 5.1%, highlighting the role of input selection. Evaluation of the U-Net baseline network with RGB + MSRCR input showed improved performance but struggled with more complex videos, causing confusion in distinguishing genuine/spoof faces. To address this, we introduced CBAM module to

focus on salient features in input, resulting in significantly improved performance. The final model, incorporating the attention mechanism, achieved a reduced ACER of 1.8%, highlighting the critical role of attention mechanisms in handling complex scenarios and enhancing the overall robustness of our proposed framework.

4.6 Visualization and Analysis

In Fig. 4, we present examples of successful visualizations produced by our framework for both live and spoof faces. The green boundary on the samples indicates live face inputs, while the red boundary denotes spoof samples. For genuine samples, our framework adeptly produces depth images that closely align with the ground truth, albeit with minor discrepancies in finer details. Conversely, for spoof faces, our framework predominantly generates zero-depth maps, occasionally resulting in images resembling noise.

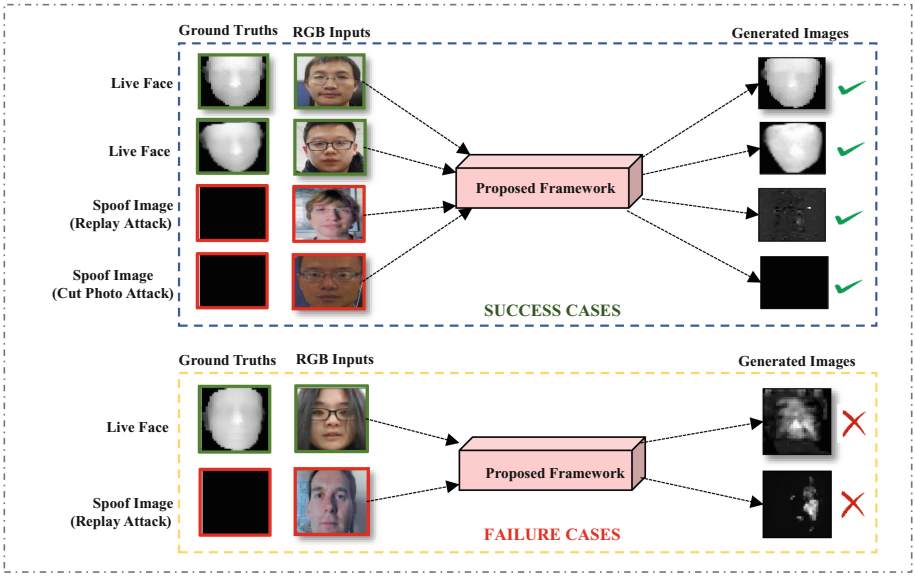


Fig. 4. Comparison of generated depth images and Ground Truths. Row1: RGB input images; Row2: Ground Truth depth images; and Row 3: Framework generated depth maps. Blue box indicates success cases, while red box display instances of failure cases. (Color figure online)

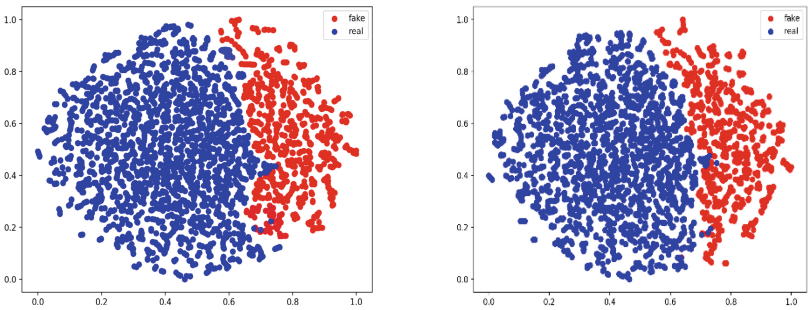


Fig. 5. The feature distribution based on t-SNE for (a) CASIA-FASD (C), (b) REPLAY-ATTACK (RA) datasets

Figure 4 also highlights some failure cases. For instance, despite being live faces, some samples processed by our framework result in inaccurate depth maps. Another example includes a spoof face where a replay video of a person is fed into the framework, resulting in patches of depth images that closely resemble facial depth maps instead of generating zero-depth maps. This highlights a fundamental challenge causing classification errors. To assess the discriminative capability of the CNN features extracted for FAS, we utilize the t-SNE visualization technique [66]. By projecting the CNN features used by our classifier, we observe distinctions between live and spoof samples, as illustrated in Fig. 5. Figure 5(a) demonstrates feature distributions for the C dataset using a model trained specifically on that dataset, while Fig. 5(b) shows distributions for the RA dataset using a model trained on RA data. These figures highlight how our proposed framework seamlessly translates RGB faces into the depth domain, resulting in comparable feature distributions for both genuine and spoofing faces across different datasets.

These observations suggest that while our framework learns a comprehensive range of features for depth map generation, it struggles to accurately capture the features that distinguish between genuine and counterfeit faces. This limitation likely stems from the use of a basic U-Net architecture for our generator, which may not fully capture the nuanced features required for precise depth map generation in all scenarios. As this is an initial experiment, there is significant scope for future work. The experimentation can be extended by enhancing the generator with advanced architectures for robust depth map generation.

5 Conclusion

In this work, we introduce an innovative GAN-based FAS approach by leveraging a GAN network for face depth map generation and extracting crucial features to discern spoof faces. The framework comprises of three integral components: a generator, a discriminator, and a classifier, employing a domain transfer process that integrates MSRCR images alongside RGB. To optimize the strengths of both inputs, we implemented the CBAM mechanism on incoming MSRCR images. This refined input is processed through the

U-Net shaped architecture to ensures the meaningful extraction of features. The decoder blocks then process these features to produce high-quality depth maps.

Our framework underwent comprehensive evaluation across three challenging databases demonstrating competitive performance in both intra- and cross-database testing scenarios. The evaluation outcomes show the proficiency of our generator in producing profound depth maps for real data and zero-depth maps for spoof samples, effectively utilizing the collaborative strengths of RGB and MSRCR data. These results emphasize the robustness and versatility of our proposed approach in addressing the intricacies of face anti-spoofing tasks across diverse datasets.

References

1. Zhang, Z., Yan, J., Liu, S., Lei, Z., Yi, D., Li, S.Z.: A face antispoofing database with diverse attacks. In: 5th IAPR ICB, New Delhi, India (2012)
2. Chingovska, I., Anjos, A., Marcel, S.: On the effectiveness of local binary patterns in face anti-spoofing. In: BIOSIG. Germany, (2012)
3. Erdogmus, N., Marcel, S.: Spoofing face recognition with 3D masks. IEEE Trans. Inf. Forensics Secur. **9**(7), 1084–1097 (2014)
4. Boulkenafet, Z., Komulainen, J., Hadid, A.: Face spoofing detection using color texture analysis. IEEE TIFS **11**(8), 1818–1830 (2016)
5. Antil, A., Dhiman, C.: MF2ShrT: multimodal feature fusion using shared layered transformer for face anti-spoofing. ACM Trans. Multimed. Comput. Commun. Appl. **20**(6), 1–21 (2024)
6. Antil, A., Dhiman, C.: A two stream face anti-spoofing framework using multi-level deep features and ELBP features. Multimedia Syst. **29**, 1361–1376 (2023)
7. Wang, Z., Xu, Y., Wu, L., Han, H., Ma, Y., Ma, G.: Multi-perspective features learning for face anti-spoofing. In: ICCVW, Montreal (2021)
8. Liu, Y., Jourabloo, A., Liu, X.: Learning deep models for face anti-spoofing: binary or auxiliary supervision. In: CVPR, UT (2018)
9. Liu, Y., Liu, X.: Spoof trace disentanglement for generic face anti-spoofing. IEEE Trans. Pattern Anal. Mach. Intell. **45**(3), 3813–3830 (2023)
10. Wu, Y., Tao, D., Luo, Y., Cheng, J., Li, X.: Covered style mining via generative adversarial networks for face anti-spoofing. Pattern Recog. **132**, 108957 (2022)
11. Wang, Y., Song, X., Xu, T., Feng, Z., Wu, X.-J.: From RGB to depth: domain transfer network for face anti-spoofing. IEEE TIFS **16**, 4280–4290 (2021)
12. Jobson, D.J., Rahman, Z., Woodell, G.A.: A multiscale retinex for bridging the gap between color images and the human observation of scenes. IEEE TIP **6**, 965–976 (1997)
13. Woo, S., Park, J., Lee, J.-Y., Kweon, I.S.: CBAM: convolutional block attention module. arXiv:1807.06521 (2018)
14. Wen, D., Han, H., Jain, A.K.: Face spoof detection with image distortion analysis. IEEE Trans. Inf. Forensics Secur. **10**(4), 746–761 (2015)
15. Boulkenafet, Z., Komulainen, J., Li, L., Feng, X., Hadid, A.: OULU-NPU: a mobile face presentation attack database with real-world variations. In: IEEE International Conference on Automatic Face & Gesture Recognition (FG) (2017)
16. Chingovska, I., Anjos, A., Marcel, S.: On the effectiveness of local binary patterns in face anti-spoofing. In: Proceedings of the International Conference of Biometrics Special Interest Group (BIOSIG), Darmstadt, Germany (2012)
17. Antil, A., Dhiman, C.: Two stream RGB-LBP based transfer learning model for face anti-spoofing. In: CVIP, India (2023)

18. Z. Yu, X. Li , X. Niu, J. Shi and G. Zhao,: Face anti-spoofing with human material perception. In: ECCV, Glasgow (2020)
19. Yang, J., Lei, Z., Li, S.Z.: Learn convolutional neural network for face anti-spoofing. In: Computer Vision and Pattern Recognition (cs.CV) (2014)
20. Lucena, O., Junior, A., Moia, V., Souza, R., Valle, E., Lotufo, R.: Transfer learning using convolutional neuralnetworks for face anti-spoofing. In: International Conference Image Analysis and Recognition, Canada (2017)
21. George, A., Marcel, S.: On the effectiveness of vision transformers for zero-shot face anti-spoofing. In: IJCB, Shenzhen, China (2021)
22. Liu, A., et al.: FM-ViT: flexible modal vision transformers for face anti-spoofing. In: IEEE TIFS (2023)
23. Chen, Z., et al.: Generalizable representation learning for mixture domain face anti-spoofing. In: AAAI, Vancouver, Canada (2021)
24. Zhou, F., et al.: Face anti-spoofing based on multi-layer domain adaptation. In: ICMEW, Shanghai, China (2019)
25. Jiang, F., Liu, Y., Si, H., Meng, J., Li, Q.: Cross-scenario unknown-aware face anti-spoofing with evidential semantic consistency learning. IEEE Trans. Inf. Forensics Secur. **19**, 3093–3108 (2024)
26. Wang, Z., Wang, Z., Yu, Z., Deng, W.: Domain generalization via shuffled style assembly for face anti-spoofing. In: CVPR, LA (2022)
27. Jia, Y., Zhang, J., Shan, S., Chen, X.: Single-side domain generalization for face anti-spoofing. In: CVPR, Seattle, WA, USA (2020)
28. Hu, C., Feng, Z.-H., Wu, X.-J., Kittler, J.: Dual encoder-decoder based generative adversarial networks for disentangled facial representation. IEEE Access **8**, 130159–130171 (2020)
29. Song, X., Chen, Y., Feng, Z.-H., Hu, G., Yu, D.-J., Wu, X.-J.: Self-growing and pruning generative adversarial networks. IEEE Trans. Neural Networks Learn. Syst. **32**, 2458–2469 (2021)
30. Ledig, C., et al.: Photo-realistic single image super-resolution using a generative adversarial network. arXiv:1609.04802 (2017)
31. Zhang, Z., Cheng, H., Li, W., Wang, P.: An improved GAN-based depth estimation network for face anti-spoofing. In: 9th ICCV, Tianjin (2023)
32. Liu, Y., Stehouwer, J., Liu, X.: On disentangling spoof trace for generic face anti-spoofing. In: ECCV, Glasgow (2020)
33. Zhang, K.-Y., et al.: Face anti-spoofing via disentangled representation learning. In: ECCV, Glasgow, United Kingdom (2020)
34. Qin, Y., et al.: Learning meta model for zero-and few-shot face anti-spoofing. In: AAAI, New York, USA (2020)
35. Liu, A., et al.: Face anti-spoofing via adversarial cross-modality translation. IEEE TIFS **16**, 2759–2772 (2021)
36. Quan, R., Wu, Y., Yu, X., Yang, Y.: Progressive transfer learning for face anti-spoofing. IEEE TIP **30**, 3946–3955 (2021)
37. Choi, D.H., Jang, I.H., Kim, M.H., Kim, N.C.: Color image enhancement based on single-scale retinex with a JND-based nonlinear filter. In: ISCAS, New Orleans, Louisiana, USA (2007)
38. Jobson, D., Rahman, Z., Woodell, G.A.: Properties and performance of a center/surround retinex. IEEE Trans. Image Process. **6**, 451–462 (1997)
39. Xie, S.J., Lu, Y., Yoon, S., Yang, J., Park, D.S.: Intensity variation normalization for finger vein recognition using guided filter based singe scale retinex. Sensors **15**(7), 17089–17105 (2015)

40. Lee, C.-H., Shih, J.-L., Lien, C.-C., Han, C.-C.: Adaptive multiscale retinex for image contrast enhancement. In: Proceedings of the 2013 International Conference on Signal-Image Technology & Internet-Based Systems, Kyoto, Japan (2013)
41. Tan, M., Le, Q.V.: EfficientNet: rethinking model scaling for convolutional neural networks. arXiv:1905.11946 (2020)
42. Isola, P., Zhu, J.-Y., Zhou, T., Efros, A.A.: Image-to-image translation with conditional adversarial networks. arXiv:1611.07004 (2018)
43. Information technology-biometric presentation attack detection—Part 1: framework. In: Document ISO/IEC JTC 1/SC 37 Biometrics, International Organization for Standardization (2016)
44. Viola, Snow: Detecting pedestrians using patterns of motion and appearance. In: ICCV, Nice (2003)
45. Zhang, K., Zhang, Z., Li, Z., Qiao, Y.: Joint face detection and alignment using multi-task cascaded convolutional networks. In: Computer Vision and Pattern Recognition, Las Vegas, NV, USA (2016)
46. Feng, Y., Wu, F., Shao, X., Wang, Y., Zhou, X.: Joint 3D face reconstruction and dense alignment with position map regression network. arXiv:1803.07835 (2018)
47. Boulkenafet, Z., Komulainen, J., Hadid, A.: Face anti-spoofing based on color texture analysis. In: ICIP, Quebec City, QC, Canada (2015)
48. Atoum, Y., Liu, Y., Jourabloo, A., Liu, X.: Face anti-spoofing using patch and depth-based CNNs. In: IJCB, Denver, CO, USA (2017)
49. Chen, H., Hu, G., Lei, Z., Chen, Y., Robertson, N.M., Li, S.Z.: Attention-based two-stream convolutional networks for face spoofing detection. IEEE Trans. Inf. Forensics Secur. **15**, 578–593 (2019)
50. Chen, H., Chen, Y., Tian, X., Jiang, R.: A cascade face spoofing detector based on face anti-spoofing R-CNN and improved retinex LBP. IEEE Access **7**, 170116–170133 (2019)
51. Wang, Y., Song, X., Xu, T., Feng, Z., Wu, X.-J.: From RGB to depth: domain transfer network for face anti-spoofing. IEEE Trans. Inf. Forensics Secur. **16**, 4280–4290 (2021)
52. Chang, H.-H., Yeh, C.-H.: Face anti-spoofing detection based on multi-scale image quality assessment. Image Vis. Comput. **121**, 104428 (2022)
53. Jingade, R.R., Kunte, R.S.: DOG-ADTCP: a new feature descriptor for protection of face identification system. Expert Syst. Appl. **201**, 117207 (2022)
54. Shu, X., Li, X., Zuo, X., Xu, D., Shi, J.: Face spoofing detection based on multi-scale color inversion dual-stream convolutional neural network. Expert Syst. Appl. **224**, 119988 (2023)
55. Z. Wang, Z. Yu, C. Zhao and X. Zhu,: Deep spatial gradient and temporal depth learning for Face Anti-Spoofing. In: CVPR, Seattle, Washington (2020)
56. Yu, Z., et al.: Searching central difference convolutional networks for face anti-spoofing. In: CVPR, Seattle, Washington (2020)
57. Yang, X., et al.: Face anti-spoofing: model matters, so does data. In: CVPR, Long Beach, CA, USA (2019)
58. Yu, Z., Li, X., Shi, J., Xia, Z., Zhao, G.: Revisiting pixel-wise supervision for face anti-spoofing. IEEE Trans. Biometrics Behav. Identity Sci. **3**, 285–295 (2021)
59. Boulkenafet, Z., et al.: A competition on generalized software-based face presentation attack detection in mobile scenarios. In: IEEE International Joint Conference on Biometrics (IJCB), Denver, CO, USA (2017)
60. Liu, Y., Jourabloo, A., Liu, X.: Learning deep models for face anti-spoofing: binary or auxiliary supervision. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA (2018)
61. Jourabloo, A., Liu, Y., Liu, X.: Face de-spoofing: anti-spoofing via noise modeling. In: ECCV, Munich, Germany (2018)

62. Pereira, T.D.F., et al.: Face liveness detection using dynamic texture. EURASIP J. Image Video Process. **2014**, 1–5 (2014)
63. Menotti, D., et al.: Deep representations for iris, face, and fingerprint spoofing detection. IEEE TIFS **10**(4), 864–879 (2015)
64. Li, H., Pan, S.J., Wang, S., Kot, A.C.: Domain generalization with adversarial feature learning. In: IEEE/CVF CVPR, USA (2018)
65. Shao, R., Lan, X., Li, J., Yuen, P.C.: Multi-adversarial discriminative deep domain generalization for face presentation attack detection. In: CVPR, Long Beach, CA (2019)
66. Maaten, L.V.D., Hinton, G.: Visualizing data using t-SNE. J. Mach. Learn. Res. **9**, 2579–2605 (2008)

Parallel Attention Based Network for Human Activity Recognition Using Wearable Devices

Chenyang Xu[1,2], Feiyi Fan[1], Guanzhou Ke[3], Changru Guo[1], Qingyu Wu[1], and Jianfei Shen[1,4(✉)]

[1] Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100090, China
shenjianfei@ict.ac.cn
[2] School of Electrical and Information Engineering, Tianjin University, Tianjin 300072, China
[3] Beijing Jiaotong University, Beijing 100044, China
[4] Jinan Zhongke Ubiquitous-Intelligent Institute of Computing Technology, Jinan 250102, Shandong, China

Abstract. In recent years, significant progress has been made in improving the efficacy of wearable human activity recognition (HAR) tasks using deep learning technology. Existing research indicates that stacked convolutional layers effectively extract high-semantic signal features from multi-sensor channel time-series data. However, these approaches disregard the fact that sensor signals are low-semantic and sensor-based deep networks lead to overfitting and gradient vanishing. In this paper, we present the parallel attention-based HAR (PA-HAR) method. Our method employs multiple small-scale receptive fields to extract low-semantic signals in parallel and a skip-squeeze excitation block to establish correlations among multi-feature maps based on the feature channel dimension. We also introduce smooth and non-monotonic sigmoid linear units (SiLU) to integrate multi-scale and cross-channel features in order to prevent the loss of non-linear information due to small-scale receptive fields and reduce representational ability loss. Extensive experiments on seven public datasets show that our proposed PA-HAR model outperforms state-of-the-art approaches in HAR tasks. In addition, we develop a wearable real-time activity recognition system based on the embedded device with our model.

Keywords: Human Activity Recognition (HAR) · Wearable Device · Parallel Attention (PA)

1 Introduction

HAR utilizing Inertial Measurement Units (IMU) is extensively applied in numerous fields due to the accelerated development of mobile sensing and ubiquitous computing. These fields include activity monitoring, healthcare, and

A. Antonacopoulos et al. (Eds.): ICPR 2024, LNCS 15313, pp. 450–465, 2025.
https://doi.org/10.1007/978-3-031-78201-5_29

human-computer interaction. HAR is a technique of automatically classifying the activities of specific people with different sensors placed in different parts of their bodies. Then, HAR utilizes algorithms to identify the activity with these sensor data.

For sensor-based HAR work, conventional Machine Learning (ML) methods, such as Random Forest (RF) and Decision Tree (DT), can be used to predict what humans are doing at any time, achieving remarkable performance. Despite the various benefits provided by conventional ML methods in HAR, they require the manual extraction of features from the raw signal data, which is usually complicated and time-consuming. Additionally, shallow features could not classify complex activities. Many deep learning methods, such as Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs), are proposed in recent years. These methods make HAR tasks more precise and efficient.

In previous studies, researchers utilize CNNs to extract the temporal and sensor channel features in signal segments (signal image) [1]. In the CNN-based HAR method, only local temporal and sensor channel features can be extracted due to limited receptive fields, ignoring long-range dependencies in the temporal dimension and cross-channel correlations in the sensor channel dimension [2]. Other researchers use 1-D filters to extract long-range temporal dependencies along this research line. Then, they utilize high-contribution channels to replace low-contribution channels by measuring the contribution of each feature channel [3]. Other scholars find that normalization operations result in small contributions from most feature channels. Thus, they reactivate feature channels by applying whitening or decorrelation operations to equalize the contribution of each channel [4]. However, these methods only reweight the feature maps in the feature channel dimension and ignore the long-range dependencies in the temporal and sensor channel dimension.

Due to the promising results achieved by RNNs in the field of time series signals, some scholars employ RNNs to extract long-range correlations in both the temporal dimension and sensor channel dimension. CNNs are then attached to fuse local temporal and sensor channel features. Unfortunately, HAR data often contains uncorrelated signals (noise), and RNNs tends to model such noise repetitively, resulting in lower classification accuracy of the model [5].

The attention mechanisms can assist in determining where to focus information while reducing repetitive or even useless information. Convolutional block attention module (CBAM) [6] is among the numerous ways of introducing attention to the HAR area [7]. CBAM attention extracts long-range dependency and cross-sensor channel correlations using a large-scale receptive field. However, large-scale receptive fields can cause feature compression [8], which reduces the resolution of sensor features and causes fine-grained feature loss for datasets with a small number of sensor channels (such as the UniMib-SHAR dataset). This is detrimental to the forward propagation of features. Other scholars utilize multi-scale receptive fields to extract signal features in parallel [9], which enables the simultaneous extraction of high-resolution and highly semantic features. However, this method only consider the multi-scale features, and ignores the correlation between multi-feature maps.

In conclusion, CNN-based approaches lose long-range dependency and cross-sensor channel correlation, whereas RNN-based methods may extract the background signal (noise) repeatedly, which is disadvantageous to activity classification. Existing attention-based HAR approaches either compress signal features or lose cross-feature channel interaction information [10].

Inspired by the pyramidal convolutional attention proposed in computer vision area [11], in this paper, we propose a parallel attention mechanism for the sensor-based HAR domain and we investigate the performance of multi-scale parallel convolutional neural networks for low-semantic sensor data. Our approach distinguishes itself from the pyramidal convolutional attention technique employed in computer vision by virtue of its lightweight design and specialized focus on extracting feature information from sensor signals. The PA mechanism for sensor-based HAR concurrently utilizes two small-scale receptive fields to extract higher semantic and high-resolution information, decreasing many signal features. Then, we propose the Skip Squeeze Excitation (SSE) to enhance the feature by establishing the correlation of the multi-feature maps. The limited non-linearity in lightweight attention may constrain the model's representational power. Hence, we replace the ReLU function with SiLU [12], which increases the non-linearity and avoids the feature loss caused by the dead zone of ReLU. The baseline network used is a 1D filter to extract only signal features, as previously indicated [4,7]. With the enhanced PA mechanism, the model can extract higher semantic features after each convolution layer, enhancing the feature space by capturing temporal-spatial information on different scales. The following is a summary of the contributions made by our work:

- We propose a parallel attention mechanism, which can further fuse features from multiple sensor channels using different scales of receptive fields. Based on the above methods, the model can extract high-resolution and high-semantic features.
- We propose a lightweight skip-squeeze excitation (SSE) block to extract global information and cross-feature channel correlation. It achieve the interaction between the multiple feature maps. In addition, we employ the SiLU activation function to introduce more nonlinear information and thus enhance the representational power of the model.
- We compare our proposed PA-HAR approach to SOTA methods on seven sensor-based datasets and show that it provides superior results. Additionally, on an embedded platform, we evaluate the practical application efficacy of the PA-HAR model, demonstrating that it is effective for HAR on lightweight mobile devices.

2 Methodology

2.1 Model Overview

In this section, we briefly introduce our proposed PA mechanism. The PA mechanism consists of three parallel substructures that process features at different

resolutions and establish the cross-feature map relationship. The input data is fed into the proposed multi-scale feature extraction block in the first substructure, which consists of multiple receptive fields. We use a 3×3 receptive field to extract the long-range correlation of temporal and sensor channel dimensions as a high semantic feature. To maintain the high-resolution sensor signal features and to avoid feature compression issues, we use a 1×1 receptive fields. Based on the above discussion, we only consider the high semantic and high resolution information, ignoring the correlation between the multi feature maps. To address this, we propose the SSE block based on the Squeeze Excitation (SE) block [13] for HAR tasks, which can establish the correlation of the multi-feature map. Compare with the SE block, the SSE block employs only a single fully connected layer, which further reduces block complexity and maintains the PA mechanism's lightweight characteristic. The PA-HAR framework is illustrated in Fig. 1.

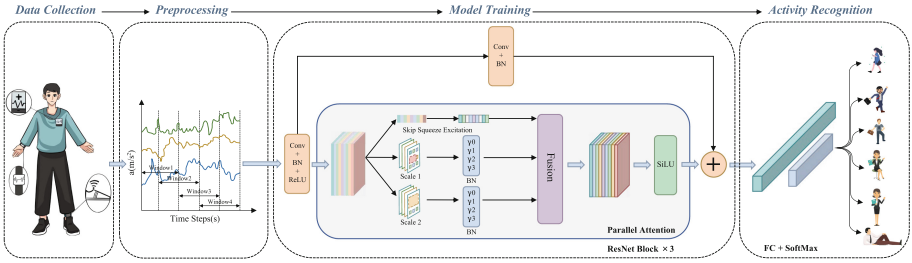


Fig. 1. Overview of Parallel Attention HAR Model Based on ResNet

2.2 Multi Scale Feature Extraction

Multiple raw sensor signals are represented as $S = \{s_1, s_2, \ldots, s_n\}$ to a predetermined window size, where $S \in \mathbb{R}^{m \times n}$, S is the signal image provided to the network, m is the duration of the time series, and n is the sensor channel dimension. Let $X \in \mathbb{R}^{C \times H \times W}$ as the signal feature map of the input network, which represents the signal image S, that is pass through a layer of the convolutional network. C represents the number of feature channels, H represents the height (temporal dimension), and W represents the width (sensor channel dimension). As shown in Fig. 1, we first utilize a multi-scale receptive field to extract the high semantic information, which embeds the long-range correlation in temporal and sensor channel dimensions. Moreover, we can maintain the high-resolution feature. All signal feature maps are fed into the multi-scale convolution block to extract the temporal and cross-sensor channel correlation. The generation function of the multi-scale feature map is as follows:

$$F_i = \text{Conv}\left(k_i \times k_i\right)(X), i = 1, 2 \tag{1}$$

where k_i represents the i-th receptive field size.

Then, we send feature maps of different scales to the batch normalization (BN) layer. The BN process can suppress less salient weights [14]. It penalizes the weight sparsity of the multi scale feature, making it more computationally efficient while maintaining equivalent performance. The following is the BN procedure:

$$Scale_i = BN\left(F_i\right) = \gamma \frac{F_i - \mu_{\mathcal{B}}}{\sqrt{\sigma_{\mathcal{B}}^2 + \epsilon}} + \beta, i = 1, 2 \tag{2}$$

where $\mu_{\mathcal{B}}$ and $\sigma_{\mathcal{B}}$ are the mean and standard deviation of mini-batch $\mathcal{B}$, respectively; γ and β are trainable affine transformation parameters (scale and shift).

2.3 Skip Squeeze Excitation Block

We aim to design a lightweight and efficient attention mechanism. To establish multi-feature graph associations, we propose the SSE block, which uses a hopping operation and only a fully connected layer compared to the SE block. The SSE feature channel extraction module consists of two parts squeeze, and excitation. In the squeeze part, we utilize the global average pooling operation to generate global feature channel information. The operation formula is as follows:

$$P = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} X(i, j) \tag{3}$$

where X is the signal feature map and P is the global pooling process.

The excitation part needs to obtain attention weights based on the above global features, computed as follows.

$$Z = \sigma(T(P)) \tag{4}$$

where T represents the fully connected layer, σ represents the Sigmoid activation function, P represents the pooling operation of all signal feature maps along the feature channel dimensions, and Z is the result generated by the SSE block. The excitation section establishes linear relationships within the feature channel dimension combinations between multiple feature maps using the fully connected layer.

2.4 Feature Fusion

As mentioned above, the lightweight attention mechanism combined with shallow networks has less non-linearity, which limits the model's representational capability, especially for dataset with multi sensors such as OPPORTUNITY [15], which include multiple complex activities. Moreover, the ReLU function has a dead zone, which causes negative gradients to be set to zero and cannot be updated. In order to increase shallow network non-linearity information, we replace the ReLU activation with SiLU. SiLU is a smooth approximation of ReLU, as is shown in Fig. 2. In the following, we derive the formula for the smooth approximation of SiLU to ReLU. For a maximum function $\max\left(m_1, m_2, \ldots, m_n\right)$, we can

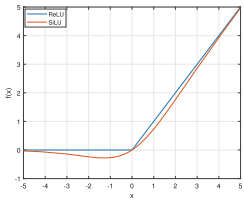


Fig. 2. Comparison of ReLU function and SiLU function.

obtain its smooth approximation by using a general approximation formula as follows:

$$F_\beta\left(m_1,\ldots,m_n\right)=\frac{\sum_{i=1}^{n} m_i e^{\beta m_i}}{\sum_{i=1}^{n} e^{\beta m_i}} \tag{5}$$

where β is introduced to control the degree of smoothing of F_β. When $\beta \to \infty$, $F_\beta \to \max$ (Nonlinear), and $\beta \to 0$, $F_\beta \to Average$ (Linear). We denote the β as 1 to close to ReLU. For $\max\left(m_1, m_2, \ldots, m_n\right)$, we consider the case when $n = 2$. Given a common activation functions are in the form of $\max\left(\eta_a(m), \eta_b(m)\right)$ (e.g. ReLU $\max(m, 0)$) where $\eta_a(m)$ and $\eta_b(m)$ denote linear functions. And we denote σ as the *Sigmoid* function and the approximation becomes:

$$\begin{aligned}
&F\left(\eta_a(m), \eta_b(m)\right)\\
&=\eta_a(m)\cdot\frac{e^{\eta_a(m)}}{e^{\eta_a(m)}+e^{\eta_b(m)}}+\eta_b(m)\cdot\frac{e^{\eta_b(m)}}{e^{\eta_a(m)}+e^{\eta_b(m)}}\\
&=\eta_a(m)\cdot\frac{1}{1+e^{-(\eta_a(m)-\eta_b(m))}}+\eta_b(m)\cdot\frac{1}{1+e^{-(\eta_b(m)-\eta_a(m))}}\\
&=\eta_a(m)\cdot\sigma\left[(\eta_a(m)-\eta_b(m))\right]+\eta_b(m)\cdot\sigma\left[(\eta_b(m)-\eta_a(m))\right]\\
&=(\eta_a(m)-\eta_b(m))\cdot\sigma\left[(\eta_a(m)-\eta_b(m))\right]+\eta_b(m)
\end{aligned} \tag{6}$$

We find that when $\eta_a(m) = m, \eta_b(m) = 0$, $\max(m, 0)$ is exactly the expression of ReLU, and $m \cdot \text{sigmoid}(m)$ is exactly the expression of SiLU. We can think of SiLU as a smooth way to get close to ReLU. We use the SiLU function to correlate multi-scale features and multi feature maps as follows:

$$Out = \text{SiLU}(Scale_1 + Scale_2 + Z) \tag{7}$$

where $Scale_1$ represents the high resolution feature, $Scale_2$ represents the high semantic feature, and Z represents the multi feature maps correlation information.

According to Fig. 2 and Eq. 6, we can see that SiLU is smooth and non-monotonic, which can introduce more nonlinear information, and the gradient is derivable. Compared to the ReLU function, the SiLU function has a smoother curve as it approaches zero and allows the network to have an output range between 0 and 1 due to the use of a Sigmoid function. The HAR data are time series data (with upper and lower bounds), better results can be achieved using SiLU than ReLU. In conclusion, SiLU is well suited for our proposed lightweight parallel attention mechanism.

2.5 Implementation

We choose three layers of the ResNet to show the benefits of the PA mechanism over earlier HAR approaches since this study aims to offer a more efficient way of boosting the convolutional features of HAR networks. ResNet's convolutional blocks comprise two convolutional layers with the same kernel size. The PA mechanism is incorporated into the ResNet network, as illustrated in Fig. 1.

3 Experiments

The benchmark datasets, experimental settings, and evaluation metrics are each presented in this section.

Table 1. The Concise Description of the Procedure for Working with the HAR Datasets.

Attribute	UCIHAR	OPPO	PAMAP2	USCHAD	UniMib	MHEALTH	DSADS
Sampling Rate	50	30	100	100	50	50	25
Train/Test	7:3	8:2	7:3	7:3	7:3	8:2	7:3
Window Size	128	64	171	512	151	100	None
Overlap Rates	50%	50%	None	50%	50%	50%	None
Categories	6	17	12	12	17	13	9

3.1 Dataset Description

We evaluate the efficacy of our model using seven widely-used HAR datasets, summarised in Table 1. The dataset processing method is set according to existing literature [1,2]. The UCIHAR Dataset [16] comprises data from 30 participants who performed 6 activities while donning a smartphone on their midsection. The PAMAP2 Dataset [17] is collected from 9 participants who wear numerous sensors, including chest, wrist, and ankle sensors, to collect data. The UniMib-SHAR Dataset [18] is collected from 30 participants using Android smartphones, with each participant must carry a smartphone in both of their front pockets. The OPPORTUNITY Dataset [15] is captured using multi-sensor modalities and peripheral sensors placed on 12 subjects. The USCHAD Dataset [19] comprises 12 physical activities and is collected from 14 volunteers. The DSADS Dataset [20] consists of data collected on 19 activities performed for 5 min by eight participants, 9 of which are utilized in our evaluation. Finally, the MHEALTH Dataset [21] contains recordings of body movements and vital signals from ten persons with varying features who completed 12 exercises.

3.2 Experimental Details

The datasets splitting strategy is summarised in Table 1. The batch size for the UniMib-SHAR dataset is 128, while the other datasets are 64. For all datasets, the beginning learning rate is 0.001, the Adam optimizer is used, and the cross-entropy function is used as the loss function. The other hyperparameters are set to their default settings.

Table 2. Model Average Accuracy (%) on Various Datasets.

Methods \ Datasets	UCIHAR	OPPO	PAMAP2	USCHAD	UniMib	MHEALTH	DSADS
Other Researchers' Results	$97.35_{\text{Huangetal.[4]}}$	$89.15_{\text{Huetal.[22]}}$	$92.14_{\text{Huangetal.[4]}}$	$91.70_{\text{Bietal.[23]}}$	$78.65_{\text{Huangetal.[4]}}$	$98.76*_{\text{Huangetal.[3]}}$	$94.44*_{\text{Huangetal.[3]}}$
	$97.38_{\text{Tangetal.[9]}}$	$87.40_{\text{Kimetal.[24]}}$	$94.29_{\text{Tangetal.[9]}}$	$91.07_{\text{Lietal.[25]}}$	$79.19_{\text{Tangetal.[9]}}$	$96.68*_{\text{Huangetal.[26]}}$	$94.52*_{\text{Huangetal.[26]}}$
	$97.23_{\text{Huangetal.[27]}}$	$81.42_{\text{Huangetal.[27]}}$	$92.25_{\text{Huangetal.[27]}}$	$85.71_{\text{Huangetal.[28]}}$	$77.52_{\text{Huangetal.[27]}}$	$90.50*_{\text{Qianetal.[29]}}$	$82.25*_{\text{Qianetal.[29]}}$
Ours	97.60	91.20	97.63	93.33	79.46	99.10	95.10
ΔSOTA	0.22 ↑	2.05↑	3.34↑	1.65↑	0.27↑	0.34↑	0.58↑

Where * represents the reproduction of results.

3.3 Comparison with Other Methods

The comparison between our approach and the SOTA approaches is shown in Table 2. The performance of the seven PA-HAR datasets significantly outperforms the SOTA (State-Of-The-Art) approaches, as shown in Table 2. Our method outperforms SOTA methods on the UCIHAR, OPPORTUNITY, PAMAP2, USCHAD, UniMib-SHAR, MHEALTH, and DSADS datasets 0.22%, 2.05%, 3.34%, 1.65%, 0.27%, 0.36%, and 0.58%, respectively.

3.4 Ablation Studies

We conduct a series of ablation experiments to demonstrate the effectiveness of the proposed PA mechanism. As seen in Fig. 3, adding PA mechanism enhances the recognition ability greatly over the baseline network. Our method improves by 1.06%, 3.24%, 2.17%, 1.35%, 2.26%, 0.84%, and 0.6% on the datasets UCI-HAR, OPPORTUNITY, PAMAP2, USCHAD, UniMib-SHAR, MHEALTH, and DSADS, respectively. The experimental results demonstrate that the PA attention mechanism can effectively extract multi-scale signal features and the correlation of multiple feature maps, which is necessary to recognize long-range periodic activities (running, cycling, etc.). These experiments demonstrate that the PA mechanism is required for HAR classification.

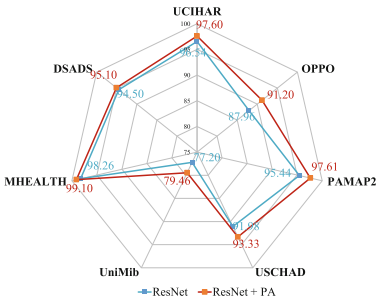


Fig. 3. Comparison of ResNet Baseline Network and ResNet Network After Adding PA Mechanism.

To evaluate the effectiveness of each part of the PA mechanism on model recognition, we conduct experiments using the UniMib-SHAR and USCHAD

datasets as examples. The results can be seen in Table 3. We conclude that enhancing the network with cross-feature maps and multi-scale feature extraction may improve classification accuracy. Experimental results demonstrate that multi-scale features are more critical for signal recognition ability because shallow convolution extracts only local dependencies and loses sensor as well as temporal correlation. We use multi-scale extraction methods to extract higher semantic features while retaining high-resolution features. The best accuracy is attained when the two components are combined.

Table 3. The Effect of Various PA Mechanism Components.

Datasets \ Network	Baseline	+SSE	+Multi Scale	+PA
UniMib-SHAR	77.20%	77.51%	79.17%	**79.46%**
USCHAD	91.98%	92.62%	93.04%	**93.33%**

3.5 Experimental Analysis

1. Channel Importance

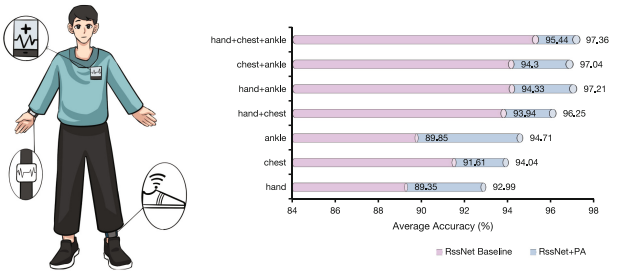


Fig. 4. The importance of sensor channels on PAMAP2 dataset.

We conduct position-dependent activity recognition studies on the PAMAP2 dataset using three IMU nodes placed at different body positions: hand, ankle, and chest. Figure 4 depicts all of the available sensor combinations and their location. We compare the baseline network with the PA-HAR network. The PA-HAR model outperforms the baseline network for various hand, chest, and ankle positions. Our findings show that a single accelerometer on the hand, ankle, or chest can accurately detect some of the most typical activities, with an average accuracy of 92.99%, 94.71%, and 94.04%, respectively. Adding the second IMU improves the classification accuracy to 96.25% (hand+chest), 97.21% (hand+ankle), and 97.04% (chest+ankle). Notably, the accuracy of activity detection is much lower for the hand and chest position than for the hand and ankle or chest and ankle positions. Adding a third IMU node yields the best classification result, suggesting that multi-modal sensor data is more beneficial for HAR.

2. PA Mechanism Position Analysis
We perform ablation experiments using the UniMib-SHAR dataset to assess the influence of the PA mechanism at various layers. The PA mechanism should be inserted after the first, second, and third layers of the ResNet model for maximum efficiency, as shown in Fig. 5. This is because ResNet extracts long-term dependence information. The PA method can extract deep cross-sensor channel association characteristics while retaining low semantic information, allowing many sensors to be correlated.

Table 4. The Effect of The Activation Function on The Model.

Datasets \ Network	Baseline	+ReLU	+SiLU
OPPORTUNITY	87.96%	89.94%	**91.20%**

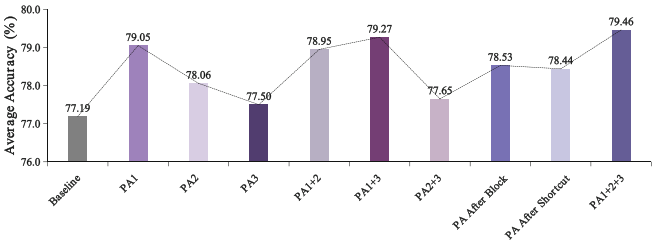


Fig. 5. Accuracy of PA Mechanism on Different Layers.

3. The Effect of the Activation Function on the Model
As shown in Table 4, using the OPPORTUNITY dataset as an example, we observe that the SiLU activation function outperforms ReLU by 1.26% in classification accuracy. Unlike ReLU, the SiLU function has a smooth and non-monotonic curve, which is differentiable at all points, which is advantageous for model optimization.

4. Time Consumption Comparison
Figure 6 shows the results of an execution time comparison for the test data of the MHEALTH dataset, which contains 1285 test data. As seen in Table 6, although our method is not the quickest, it has no apparent downsides. Our method is slower in terms of execution. Our PA mechanism increases the multi-scale convolution and pooling processes, prolonging execution time. Several superior results, including the Local-Loss method [30], solely rely on cosine similarity to calculate local losses in front of the network. To extract the largest mean difference, DDNN [29] employs a full connection mapping to extract the maximum mean difference to high-dimensional space.

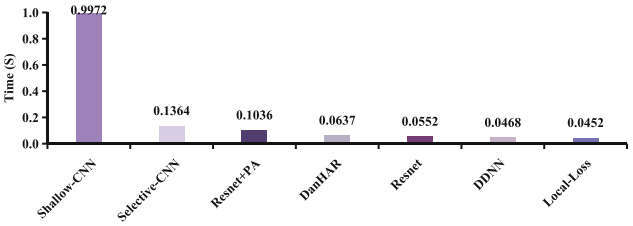


Fig. 6. The Execution Time Comparison on Different Methods

5. Compare with Other Computer Vision Attention Mechanisms

We embed four attention mechanisms (such as SE [13], CBAM [6], EPSA [11], and NAM [14]) in the area of computer vision into our baseline network. We compare them with our method to demonstrate the importance of the proposed parallel attention mechanism for HAR classification tasks.

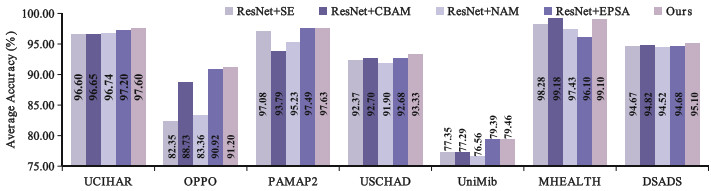
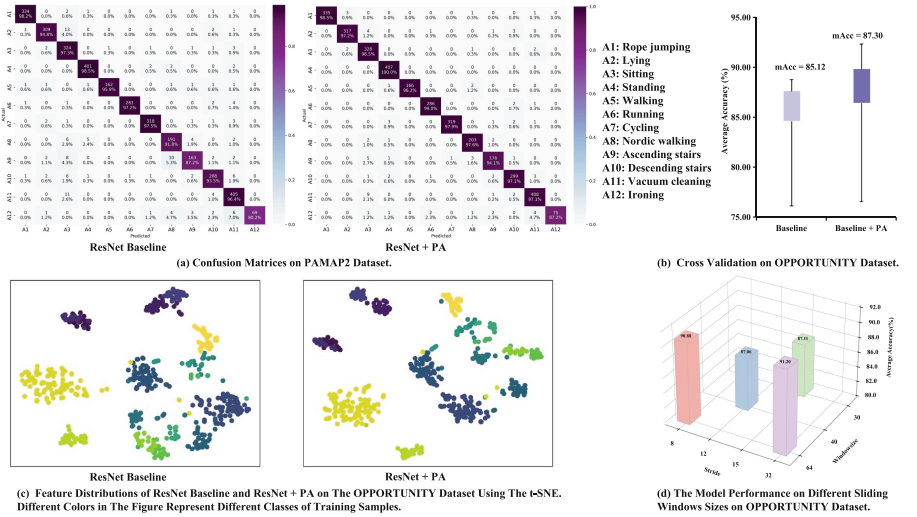


Fig. 7. Accuracy on Other Computer Vision Attention Mechanisms

As is shown in Fig. 7, we can see that our method has excellent performance on six datasets. We find that the model based on the SE attention mechanism performs poorly on OPPORTUNITY because the dataset belongs to complex activities. The SE mechanism uses pooling operations, which lose spatial-temporal correlation information, and reduce the downstream classification accuracy. The CBAM-based attention mechanism method improves classification accuracy compared with the SE-based method. However, for the UniMib-SHAR dataset, which uses a small number of sensor channels, the large-scale receptive field causes feature compression and reduces the resolution of feature maps. To solve the problem of feature compression caused by large-scale receptive fields, Zhang et al. [11] propose a multi-scale feature extraction attention mechanism (EPSA). We apply this method to the HAR domain. Compared with CBAM-based methods, this method has an improvement on multiple datasets. However, this method uses four receptive parallel fields, which increases the model's complexity. Moreover, the EPSA mechanism employs a 9×9 receptive field to extract features, which is unsuitable for HAR data with low semantic content.



(a) Confusion Matrices on PAMAP2 Dataset.

(b) Cross Validation on OPPORTUNITY Dataset.

(c) Feature Distributions of ResNet Baseline and ResNet + PA on The OPPORTUNITY Dataset Using The t-SNE. Different Colors in The Figure Represent Different Classes of Training Samples.

(d) The Model Performance on Different Sliding Windows Sizes on OPPORTUNITY Dataset.

Fig. 8. Discussion and Analysis of Experiments

Compared with the PA-HAR method, the classification accuracy of the EPSA-based methods is lower. On the MHEALTH dataset, our PA-HAR method did not outperform the CBAM-HAR method. We consider this because this dataset has multiple periodic activities, such as cycling. We use only two low-scale receptive fields to lose some long-range dependency features. The CBAM-based method can extract relatively long-range association information by the 7×7 receptive field [6]. However, compared to the CBAM-based method, our approach reduces it by only 0.08%. The EPSA-based method performs poorly on the MHEALTH dataset because it employs receptive fields of 7×7 and 9×9 sizes. While it can extract long-range dependency correlations, it also compresses features with low semantic signals. Additionally, too large a receptive field may extract boundary activity signals, such as the transition from "cycling" to "standing". These findings confirm that the PA-HAR method is robust across different datasets.

6. Performance on Cross Validation

On the OPPORTUNITY dataset, we use cross-validation to test the robustness of our approach. Unlike existing methods [2] that only use four specific data files as the test set, we cross validated all combinations of all files in the dataset as the validation set. As shown in Fig. 8(c), the PA-HAR block consistently produces a performance gain.

7. Effect of Sliding Window Size on the Model

We investigate the influence of the sliding window length. As shown in Fig. 8 (d), a smaller sliding window typically results in poor recognition accuracy. We find that using a window size of 64 and a stride of 32 gives the best performance for the OPPORTUNITY dataset.

3.6 Discussion

According to existing research [31,32] indicated, we demonstrate the benefits of ResNet in classification using confusion matrices as shown in Fig. 8 (a), the confusion matrices of the suggested model and the baseline ResNet for the HAR task on the PAMAP2 dataset. It is clear that the PA-HAR model has fewer misclassifications when compared to the baseline ResNet for two similar activities, namely "Ascending Stairs" (A9) and "Descending Stairs" (A10). We consider that the PA mechanism provides high-semantic sensor channel correlation characteristics, considerably improving this activity's classification accuracy.

We present t-SNE diagrams of the baseline network and the PA-HAR model on the OPPORTUNITY dataset in Fig. 8(c). The figure show lower intra-class distances between features within the same class. For instance, Fig. 8(d) shows that it is difficult to distinguish between the green and yellow classes, but Fig. 8(c) shows that the characteristics produced by PA-HAR are more discriminative. Based on the visualization results, it is further demonstrated that our PA-HAR model, which extracts higher semantic features and preserves low-resolution features, also establishes the association of multi-feature maps. These features are significant for our activity signal classification.

3.7 Online Prediction System

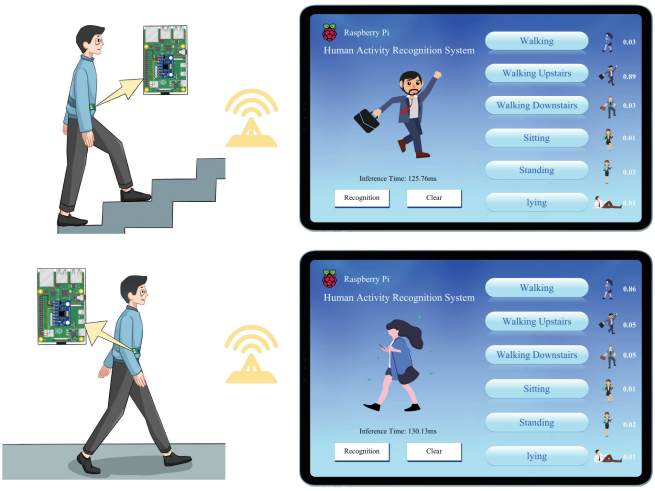


Fig. 9. Online HAR System on Raspberry Pi 4B Platform

We deploy the PA-HAR model on the Raspberry Pi platform to verify the practical effectiveness of the PA-HAR method for online inference. We use the UCI-HAR dataset as the training data and place the Raspberry Pi and IMU in the

same position as the UCIHAR collect data to ensure the online inference effect. The embedded development platform is Raspberry Pi 4B with 2 GB memory, and the IMU is MPU9250. The inference results are transmitted to the display via WiFi. Figure 9 displays the online reasoning interface for activity recognition. The Raspberry Pi accurately predicts the activity as standing when the actual activity is "Walking Upstairs", with a confidence of 89% from the softmax classifier and an inference time of 125.76 ms.

4 Conclusion

This paper proposes a HAR framework named PA-HAR, which includes a parallel attention mechanism and a ResNet network. The parallel attention mechanism effectively extracts the multi-scale feature and multi-feature map correlation. Then, we introduce non-linearity information to enhance the network's representational power. The parallel attention mechanism is embedded into a three-layer ResNet network for HAR classification. The results of experiments on seven HAR datasets show that the method we proposed works better. Additionally, plenty of ablation experiments and interpretative analyses show that each part of the structure works.

Acknowledgements. This study is supported by the Strategic Priority Research Program of Chinese Academy of Sciences, Grant No. XDA28040500; National Natural Science Foundation of China, Grant No. 62101530; Jinan Social and People's Livelihood Major Project No. 20231701.

References

1. Xu, C., Mao, Z., Fan, F., Qiu, T., Shen, J., Gu, Y.: A shallow convolution network based contextual attention for human activity recognition. In: Longfei, S., Bodhi, P. (eds.) International Conference on Mobile and Ubiquitous Systems: Computing, Networking, and Services, pp. 155–171. Springer, Cham (2022). https://doi.org/10.1007/978-3-031-34776-4_9
2. Xu, C., Shen, J., Fan, F., Qiu, T., Mao, Z.: An enhanced human activity recognition algorithm with positional attention. In: Asian Conference on Machine Learning, pp. 1181–1196. PMLR (2023)
3. Huang, W., Zhang, L., Teng, Q., Song, C., He, J.: The convolutional neural networks training with channel-selectivity for human activity recognition based on sensors. IEEE J. Biomed. Health Inform. **25**(10), 3834–3843 (2021)
4. Huang, W., Zhang, L., Wu, H., Min, F., Song, A.: Channel-equalization-har: a light-weight convolutional neural network for wearable sensor based human activity recognition. IEEE Trans. Mobile Comput. **22**, 5064–5077 (2022)
5. Zeng, M., et al.: Understanding and improving recurrent networks for human activity recognition by continuous attention. In: Proceedings of the 2018 ACM International Symposium on Wearable Computers, pp. 56–63 (2018)
6. Woo, S., Park, J., Lee, J.-Y., Kweon, I.S.: CBAM: convolutional block attention module. In: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (eds.) ECCV 2018. LNCS, vol. 11211, pp. 3–19. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-01234-2_1

7. Gao, W., Zhang, L., Teng, Q., He, J., Hao, W.: Danhar: dual attention network for multimodal human activity recognition using wearable sensors. Appl. Soft Comput. **111**, 107728 (2021)
8. Hou, Q., Zhou, D., Feng, J.: Coordinate attention for efficient mobile network design. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 13713–13722 (2021)
9. Tang, Y., Zhang, L., Min, F., He, J.: Multiscale deep feature learning for human activity recognition using wearable sensors. IEEE Trans. Industr. Electron. **70**(2), 2106–2116 (2022)
10. Tang, Y., Zhang, L., Hao, W., He, J., Song, A.: Dual-branch interactive networks on multichannel time series for human activity recognition. IEEE J. Biomed. Health Inform. **26**(10), 5223–5234 (2022)
11. Zhang, H., Zu, K., Lu, J., Zou, Y., Meng, D.: EPSANet: an efficient pyramid squeeze attention block on convolutional neural network. In: Proceedings of the Asian Conference on Computer Vision, pp. 1161–1177 (2022)
12. Ramachandran, P., Zoph, B., Le, Q.V.: Searching for activation functions. arXiv preprint arXiv:1710.05941 (2017)
13. 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)
14. Shao, Z., Hoffmann, N., et al.: NAM: normalization-based attention module. In: NeurIPS 2021 Workshop on ImageNet: Past, Present, and Future (2021)
15. Roggen, D., et al.: Collecting complex activity datasets in highly rich networked sensor environments. In: 2010 Seventh International Conference on Networked Sensing Systems (INSS), pp. 233–240. IEEE (2010)
16. Anguita, D., Ghio, A., Oneto, L., Parra, X., Reyes-Ortiz, J.L.: Human activity recognition on smartphones using a multiclass hardware-friendly support vector machine. In: Proceedings of the 4th International Conference on Ambient Assisted Living and Home Care (2012)
17. Reiss, A., Stricker, D.: Introducing a new benchmarked dataset for activity monitoring. In: 2012 16th International Symposium on Wearable Computers, pp. 108–109. IEEE (2012)
18. Micucci, D., Mobilio, M., Napoletano, P.: UniMiB SHAR: a dataset for human activity recognition using acceleration data from smartphones. Appl. Sci. **7**(10), 1101 (2017)
19. Mi, Z., Sawchuk, A.A.: USC-HAD: a daily activity dataset for ubiquitous activity recognition using wearable sensors. In: UbiComp '12: Proceedings of the 2012 ACM Conference on Ubiquitous Computing (2012)
20. Altun, K., Barshan, B., Tunel, O.: Comparative study on classifying human activities with miniature inertial and magnetic sensors. Pattern Recogn. **43**(10), 3605–3620 (2010)
21. Banos, O., et al.: Design, implementation and validation of a novel open framework for agile development of mobile health applications. Biomed. Eng. Online **14**, 1–20 (2015)
22. Chunyu, H., Chen, Y., Lisha, H., Peng, X.: A novel random forests based class incremental learning method for activity recognition. Pattern Recogn. **78**, 277–290 (2018)
23. Bi, H., Perello-Nieto, M., Santos-Rodriguez, R., Flach, P.: Human activity recognition based on dynamic active learning. IEEE J. Biomed. Health Inform. **25**(4), 922–934 (2020)

24. Kim, E.: Interpretable and accurate convolutional neural networks for human activity recognition. IEEE Trans. Industr. Inf. **16**(11), 7190–7198 (2020)
25. Li, C., Niu, D., Jiang, B., Zuo, X., Yang, J.: Meta-HAR: federated representation learning for human activity recognition. In: Proceedings of the Web Conference 2021, pp. 912–922 (2021)
26. Huang, W., Zhang, L., Gao, W., Min, F., He, J.: Shallow convolutional neural networks for human activity recognition using wearable sensors. IEEE Trans. Instrum. Meas. **70**, 1–11 (2021)
27. Huang, W., Zhang, L., Wang, S., Hao, W., Song, A.: Deep ensemble learning for human activity recognition using wearable sensors via filter activation. ACM Trans. Embed. Comput. Syst. **22**(1), 1–23 (2022)
28. Deldari, S., Smith, D.V., Xue, H., Salim, F.D.: Time series change point detection with self-supervised contrastive predictive coding. In: Proceedings of the Web Conference 2021, pp. 3124–3135 (2021)
29. Qian, H., Pan, S.J., Da, B., Miao, C.: A novel distribution-embedded neural network for sensor-based activity recognition. In: IJCAI, vol. 2019, pp. 5614–5620 (2019)
30. Teng, Q., Wang, K., Zhang, L., He, J.: The layer-wise training convolutional neural networks using local loss for sensor-based human activity recognition. IEEE Sens. J. **20**(13), 7265–7274 (2020)
31. Guo, C., Zhang, Y., Chen, Y., Xu, C., Wang, Z.: Modality consistency-guided contrastive learning for wearable-based human activity recognition. IEEE Internet Things J. **11**, 21750–21762 (2024)
32. Xu, C., Fan, F., Shen, J., Wang, H., Zhang, Z., Meng, Q.: An EEG-based depressive detection network with adaptive feature learning and channel activation. In: Proceedings of the Annual Meeting of the Cognitive Science Society, vol. 46 (2024)

Author Index

A. Antonacopoulos et al. (Eds.): ICPR 2024, LNCS 15313, pp. 467–468, 2025.
https://doi.org/10.1007/978-3-031-78201-5